POLITECHNIKA WARSZAWSKA

DYSCYPLINA NAUKOWA INŻYNIERIA LĄDOWA, GEODEZJA i TRANSPORT

DZIEDZINA NAUK INŻYNIERYJNO-TECHNICZNYCH

Rozprawa doktorska

mgr inż. Paulina Zachar

Wykorzystanie sztucznej inteligencji do wykrywania obiektów na danych fotogrametrycznych

Promotor

prof. dr hab. inż. Zdzisław Kurczyński

WARSZAWA 2024

Podziękowania

Wojciechowi Ostrowskiemu,

za opiekę merytoryczną, udzielone
wsparcie, nieocenioną pomoc,
cenne uwagi i osobiste zaangażowanie.

Łukaszowi Wilkowi,

za wsparcie informatyczne i dobre rady.

Mężowi,

za motywację, wyrozumiałość,
cierpliwość i nieustające wsparcie.

Streszczenie

Rozpowszechnienie sztucznej inteligencji, w szczególności metod głębokiego uczenia znacząco wpłynął na rozwój fotogrametrii i teledetekcji, a wykrywanie obiektów na zdjęciach lotniczych i satelitarnych stało się kluczowym obszarem badań w zakresie obserwacji Ziemi. Celem niniejszej rozprawy doktorskiej jest opracowanie metodyki wykrywania obiektów na ukośnych zdjęciach lotniczych z wykorzystaniem metod głębokiego uczenia, a dokładniej konwolucyjnych sieci neuronowych w celu zasilania miejskich baz danych. Przyjęta metodyka oparta jest na badaniach, w tym eksperymentach przeprowadzonych na różnych zbiorach danych oraz weryfikacji zaproponowanych rozwiązań i podejść celem odpowiedzi na kolejne postawione pytania badawcze. Wyniki analiz i eksperymentów pozwoliły wykazać możliwości zastosowania metod głębokiego uczenia w wykrywaniu obiektów miejskich z wykorzystaniem lotniczych zdjęć ukośnych, a także określenia lokalizacji w terenowym układzie współrzędnych i automatyzacji zasilania baz danych obiektów w przestrzeni miejskiej.

Słowa kluczowe: uczenie maszynowe, wykrywanie obiektów, ukośne zdjęcia lotnicze, dane fotogrametryczne, zbiory uczące

Abstract

The prevalence of artificial intelligence, in particular deep learning methods, has significantly influenced the development of photogrammetry and remote sensing, and object detection in aerial and satellite images has become a crucial area of research in Earth observation. This dissertation aims to develop a methodology for detecting objects in oblique aerial images using deep learning methods and, more specifically, convolutional neural networks to supply urban databases. The methodology adopted is based on research, including experiments conducted on different datasets and verification of the proposed solutions and approaches to answer the subsequent research questions posed. The results of the analyses and experiments made it possible to demonstrate the applicability of deep learning methods in the detection of urban objects using oblique aerial photographs, as well as the determination of location in the terrain coordinate system and the automation of the creation of databases of objects in urban space.

Keywords: machine learning, object detection, oblique aerial images, photogrammetric data, training datasets

Spis treści

Wstęp

Część teoretyczna

Część eksperymentalna

WSTĘP

• Opis problemu badawczego

W dzisiejszym świecie, gdzie dynamiczny rozwój technologiczny widoczny jest praktycznie w każdej dziedzinie, potrzeba szybkiego dostępu do aktualnych danych przestrzennych o dużej dokładności jest niezwykle wysoka. Rozpowszechnianie się metod sztucznej inteligencji (*ang. artificial intelligence, AI*), w szczególności metod głębokiego uczenia maszynowego (*ang. deep learning, DL*) poskutkowało szerokim zastosowaniem tych technik i algorytmów w pozyskiwaniu informacji i wydobywaniu wiedzy. Postęp widzenia komputerowego odbił się również szerokim echem w dziedzinie fotogrametrii i teledetekcji, zwłaszcza w obszarach związanych z rozpoznawaniem obrazów, klasyfikacją czy wykrywaniem obiektów.

Zwiększenie wydajności procesorów graficznych przyczyniło się do tego, że algorytmy głębokiego uczenia w ostatnich latach osiągnęły wysoką skuteczność w detekcji obiektów. Co więcej, zainteresowanie tematyką ma cały czas tendencję wzrostową i wciąż powstają nowe modele, a poprzednie wersje są ulepszane, otwierając nowe możliwości.

Wspomniany wzrost możliwości wykorzystania sztucznej inteligencji, w tym głębokiego uczenia maszynowego i konwolucyjnych sieci neuronowych spowodował, że temat detekcji obiektów z wykorzystaniem danych fotogrametrycznych pozyskiwanych z różnych pułapów (w tym zobrazowań lotniczych, satelitarnych, zdjęć z drona i chmur punktów) zyskał na popularności. Do tej pory powstało wiele różnych rozwiązań, dzięki którym detekcja obiektów na obrazie jest możliwa. Zaczynając od klasycznych metod, które opierają się na detektorach krawędzi czy klasyfikatorach bazujących na cechach, aż po złożone modele AI.

Konsekwencją tworzenia nowych modeli CNNs (*ang. convolutional neural networks*) było naturalnie zwiększenie potrzeby na nowe różnorodne zbiory danych treningowych. Jak wiadomo precyzyjna detekcja z wykorzystaniem modeli głębokiego uczenia wymaga wielu przykładów danych treningowych (*ang. ground truth annotations*), a proces ten wiąże się najczęściej z manualną pracą i oznaczaniem każdego zdjęcia po kolei, a co za tym idzie proces ten jest czasochłonny i kosztowny. Do tej pory oczywiście wiele zespołów poświęciło dużo czasu na tworzeniu takich zestawów danych, co zaowocowało udostępnieniem do użytku publicznego takich baz, jak ImageNet, PASCAL VOC, czy MS COCO. Jednakże te największe zestawy danych uczących składają się z naturalnych scen naziemnych. Chociaż można

stosować je do wstępnego trenowania modeli, nadal istnieją pewne ograniczenia, które nie pozwalają ominąć w pełni ręcznej adnotacji.

Oznaką zaawansowania technologicznego jest również nieustanne polepszanie sensorów, które umożliwiają zbieranie danych o wysokiej rozdzielczości, a co za tym idzie precyzyjne odwzorowanie terenu. Możliwość pozyskiwania danych obrazowych o wysokiej rozdzielczości w krótkim czasie poskutkowało rozwojem algorytmów sztucznej inteligencji, w szczególności modeli głębokiego uczenia do rozpoznawania obiektów na zobrazowaniach lotniczych, scenach satelitarnych, zdjęciach UAV (*ang. unmanned aerial system*), a także ortofotomapach.

W przeciągu ostatnich 20 lat zespoły badawcze udostępniły publicznie zbiory danych do detekcji na zobrazowaniach z pułapu lotniczego czy satelitarnego, wśród których dwa największe to DIOR, który zawiera 23 463 obrazy, obejmujące 20 klas obiektów oraz DOTA, który składa się z 15 kategorii obiektów i 2806 zdjęć. Niemniej jednak zestawy te nie są tak liczne, jak wspominane zbiory składające się ze zdjęć naziemnych. Ponadto różnorodność kategorii obiektów jest również niewielka, a najczęściej spotykane klasy obejmują takie obiekty, jak samochody, statki, czy samoloty. Co więcej samo zadanie detekcji obiektów na zdjęciach lotniczych jest dużo trudniejsze niż w przypadku tradycyjnej detekcji na scenach naturalnych, a do głównych wyzwań zalicza się zmienność skali, orientacji i kształtu obiektów na powierzchni Ziemi.

Pomimo wspomnianych wyzwań dzięki wykorzystaniu danych z sensorów lotniczych można w łatwy sposób uzyskać informacje o lokalizacji obiektu w terenowym układzie odniesienia, gdyż dane przestrzenne mają zazwyczaj nadaną georeferencję. Tak więc możliwości aplikacyjne wyników detekcji z wykorzystaniem zdjęć lotniczych dzięki informacji przestrzennej są dużo wyższe niż w przypadku detekcji na scenach naturalnych.

Jak zostało nakreślone, dotychczasowe badania koncentrują się głównie na scenach naturalnych oraz nadirowych zdjęciach lotniczych i satelitarnych. O ile tematyka detekcji z wykorzystaniem konwolucyjnych sieci neuronowych jest stosunkowo rozwinięta jeśli chodzi o tego typu dane, to obszarem, którego potencjał nie jest jeszcze w pełni wykorzystany jest fotogrametria bazująca na zdjęciach ukośnych. Zdjęcia ukośne stanowią źródło danych, gdzie obiekty są fotografowane z różnych perspektyw w różnej skali i mogą być odwzorowane kilkukrotnie na zdjęciach wykonanych z różnych kierunków w odstępach czasowych, co dodatkowo zwiększa rozmiar i różnorodność zbioru danych. Ponadto, w przeciwieństwie do zdjęć nadirowych, zdjęcia ukośne ukazują więcej szczegółów o obiektach, dlatego też widoczne jest

rosnące zainteresowanie tego typu zobrazowaniami, szczególnie dla obszarów miejskich. Jednakże dostępność zbiorów uczących w odniesieniu do ukośnych zdjęć lotniczych jest nadal niska. Zaczynają pojawiać się prace, gdzie wykorzystuje się na przykład fotogrametrię UAV, która jest opłacalnym podejściem do pozyskiwania zdjęć ukośnych, szczególnie dla mniejszych obszarów, ale w przypadku tej technologii problem braku danych uczących również stanowi główną barierę w badaniach.

Intensyfikacja wykorzystania metod głębokiego uczenia do detekcji obiektów na zdjęciach oraz nakreślony problem badawczy związany z brakiem zbiorów treningowych poskutkował tym, że zaczęły rozwijać się również rozwiązania wspomagające generowanie danych uczących oraz metody automatyzacji procesów tworzenia takich zbiorów. Podejścia oparte o transfer wiedzy (*ang. transfer learning*), czy wykorzystanie metod augmentacji danych (*ang. data augmentation*) to przykładowe rozwiązania.

Jednym z podejść, gdzie wykorzystywane jest przechodzenie między reprezentacjami tego samego obiektu z wykorzystaniem różnych typów danych jest metodyka zaproponowana przez Laupheimera i Haalę (2021). Polega na transformacji etykiet z ręcznie adnotowanych chmur punktów na model siatkowy (*ang. mesh*), a następnie do układu pikselowego obrazu w celu realizacji semantycznej analizy sceny. Innym przykładem jest opisana w pracy (Zachar i in., 2022) metodyka półautomatycznego tworzenia szkoleniowych zbiorów danych do wykrywania na nadirowych i ukośnych zdjęciach lotniczych z wykorzystaniem chmur punktów.

Problem badawczy poruszony w ramach prac realizowanych podczas doktoratu nawiązuje ściśle do tematyki wykorzystania metod i narzędzi sztucznej inteligencji do wykrywania obiektów na danych fotogrametrycznych. Dlatego też wszystkie wymienione aspekty powiązane z tematyką zostały poruszone w kolejnych rozdziałach, a zadanie wykrywania odniesione zostało do wybranych obiektów.

Analiza przestrzeni miejskiej cieszy się coraz większym zainteresowaniem wielu środowisk badawczych, co jest głównie związane z przyspieszoną urbanizacją i ze zwiększaniem liczby ludności zamieszkującej tereny zurbanizowane. Według Organizacji Narodów Zjednoczonych, do 2050 roku ok. 68% populacji ma zamieszkiwać miasta. W związku z czym pojawiają się nowe wyzwania w zarządzaniu przestrzenią miejską, a wykrywanie obiektów i zasilanie baz danych informacjami o wykrytych obiektach jest szczególnie ważne dla rozwoju inteligentnych miast. Dlatego też zdecydowano się na koncentrację badań wokół obiektów małej architektury miejskiej.

Istotną rolę w przestrzeni miejskiej odgrywają obiekty takie, jak ławki, fontanny, czy obiekty drogowe - latarnie uliczne, znaki drogowe, sygnalizacja świetlna. Obiekty te są istotne nie tylko z punktu widzenia zarządzania miastem, ale także planowania przestrzennego (Jin i in., 2014). Ponadto nieodłączny element krajobrazu miejskiego stanowią budynki i ich elementy takie jak fasady, okna, balkony, drzwi. Rozmieszczenie i charakterystyka wymienionych elementów mogą dostarczyć informacji na temat struktury budynków oraz ogólnego układu urbanistycznego. Dlatego też znajomość lokalizacji takich obiektów w układzie współrzędnych terenowych ma kluczowe znaczenie. Biorąc pod uwagę złożoność problemu wykrywania różnego typu obiektów zdecydowano się na ograniczenie do dwóch kategorii obiektów o różnej charakterystyce. Ostatecznie analiza wykonalności zadania detekcji na zobrazowaniach ukośnych została przeprowadzona dla latarni ulicznych i okien. Wybrane obiekty są dobrą reprezentacją obiektów małej architektury miejskiej, a problemy metodyczne wykrywania i pomiaru są bardzo zbliżone np. do banerów reklamowych, balkonów w przypadku okien czy do słupów trakcyjnych, znaków drogowych w przypadku latarni. Można więc uznać, że postawione i rozwiązane problemy badawcze odniesione do okien i latarni ulicznych są reprezentacyjne dla szerszego zbioru klas obiektów małej architektury.

Do głównych źródeł motywacji można zaliczyć przede wszystkim nieustanny rozwój metod głębokiego uczenia, wykorzystanie potencjału lotniczych zdjęć ukośnych, a także automatyzację zasilania baz danych obiektów przestrzeni miejskiej. Przyjęta metodyka zaś zakłada podejmowanie prób odpowiedzi na kolejne postawione pytania badawcze przeprowadzając szereg eksperymentów. W ramach przyjętej metodyki zakłada się wykorzystanie metod analizy danych ilościowych, jak i jakościowych, a także porównywanie i analizę zaproponowanych metod i podejść.

- ## Cel i pytania badawcze

Główny cel to opracowanie metodyki wykrywania obiektów na ukośnych zdjęciach lotniczych z wykorzystaniem metod głębokiego uczenia, a dokładniej konwolucyjnych sieci neuronowych w celu zasilania miejskich baz danych.

Do celów szczegółowych można zaliczyć:

- ✓ Wybór klas obiektów kluczowych z perspektywy zarządzania przestrzenią miejską;
- ✓ Dobór danych fotogrametrycznych do zadania detekcji obiektów małej architektury miejskiej;
- ✓ Rozwiązanie problemu braku danych uczących;
- ✓ Porównanie różnych metod detekcji obiektów;
- ✓ Wykorzystanie aspektu przestrzennego – zasilanie baz obiektów miejskich;
- ✓ Ocena metod określania położenia obiektów w układzie terenowym.

Zdefiniowane pytania badawcze brzmią następująco:

1. Jak automatyzować proces tworzenia zbiorów uczących o dużej dokładności?
2. Jakie rozwiązania głębokiego uczenia wykazują najwyższe dokładności w zadaniu detekcji obiektów na ukośnych zdjęciach lotniczych?
3. Czy zasadne jest wykorzystywanie metod sztucznego poszerzania zbiorów uczących (*ang. data augmentation*) oraz transferu wiedzy w celu uzyskania wyników detekcji o dużej dokładności?
4. Która metoda wyznaczania lokalizacji obiektów wykrytych na zdjęciach daje najlepsze rezultaty dla danej klasy obiektów?

- ## Teza pracy

Zastosowanie metod głębokiego uczenia wykazuje przydatność w wykrywaniu obiektów miejskich, a wykorzystanie potencjału lotniczych zdjęć ukośnych umożliwia określenie położenia obiektów w terenowym układzie współrzędnych i może być źródłem zasilania baz danych o obiektach miejskich.

- ## Metodyka badań

Jak podkreślano we wstępie, wzrost liczby modeli do detekcji obiektów na zdjęciach, które osiągają naprawdę dobre wyniki był jedną z motywacji badań. Jednakże celem niniejszej pracy nie jest opracowanie jak najlepszego modelu do wykrywania obiektów na lotniczych zdjęciach ukośnych, a ukazanie problemów badawczych, które mogą pojawić się w całym ciągu technologicznym jeśli chodzi o to zadanie. Zaczynając od problemu jakie dane fotogrametryczne są wystarczająco dokładne, aby móc z ich wykorzystaniem wykrywać konkretne obiekty. Następnie szeroko poruszona będzie problematyka braku zbiorów uczących wraz z przedstawieniem propozycji automatyzacji procesu tworzenia takich zbiorów. Kolejna część to analiza różnych rozwiązań oraz modeli do rozpoznawania obiektów, a końcowe wyzwanie to odpowiedź na pytanie jak wykorzystać wyniki detekcji na zdjęciach.

Pierwsza część badań dotyczyć będzie zatem doboru najlepszego źródła danych oraz oceny ich jakości i dokładności pod kątem wykorzystania do wykrywania latarni ulicznych i okien, które zostały wybrane świadomie jako obiekty odznaczające się dużą reprezentatywnością wśród innych klas obiektów przestrzeni miejskiej. Biorąc pod uwagę charakterystykę wybranych obiektów miejskich i właściwości różnych danych fotogrametrycznych, konieczne jest rozważenie, jakie podejście do ich detekcji będzie najlepsze oraz jakiego źródła danych użyć, aby obiekty były w pełni odwzorowane. W przyjętej metodyce pożądanymi danymi testowymi są ukośne zdjęcia lotnicze o odpowiedniej rozdzielczości i szczegółowości. Weryfikacja tej tezy powinna być oczywiście poparta przeglądem literatury, który będzie zawarty w pracy. Natomiast wykonane eksperymenty powinny pozwolić potwierdzić założenia dotyczące doboru zdjęć ukośnych jako najlepszego źródło danych do wykrywania wybranych obiektów w przestrzeni miejskiej.

Kolejny aspekt badawczy będzie dotyczył problemu braku danych uczących i automatyzacji tego procesu. Założeniem metodyki jest wykorzystanie różnych źródeł danych do automatycznego tworzenia zbiorów danych do szkolenia modeli detekcji obiektów. Konieczne będzie zatem zweryfikowanie kompletności i dokładności tych danych. Zbadanie potencjału wykorzystania chmur punktów jako wsparcia przy tworzeniu szkoleniowego zbioru danych na ukośnych zdjęciach lotniczych powinno odpowiedzieć na pytanie czy zaproponowane podejście pozwoli automatyzować proces tworzenia zbiorów uczących o dużej dokładności.

Chociaż stosowanie różnych metod do wykrywania obiektów miejskich na różnych typach danych geoprzestrzennych jest możliwe, zadanie to jest dość trudne. Pojawia się zatem pytanie jakie rozwiązanie będzie najlepszym wyborem. Metodyka przyjęta w pracy zakłada detekcję wybranych obiektów z wykorzystaniem modeli głębokiego uczenia. W przypadku latarni ulicznych realizowane będzie zadanie detekcji, zaś dla okien segmentacja. Ta część odnosi się zatem do odpowiedzi na pytania jakie modele głębokiego uczenia wykazują najwyższe dokładności w zadaniu wykrywania obiektów na ukośnych zdjęciach lotniczych. W ramach eksperymentów zweryfikowane zostaną również założenia dotyczące wykorzystania metod sztucznego poszerzania zbiorów uczących oraz transferu wiedzy, których celem jest poprawa dokładności i wydajności zastosowanych algorytmów.

Ostatnia część metodyki odnosić się będzie do możliwości geolokalizacji rozpoznanych obiektów. Założeniem jest wykazanie możliwości aplikacyjnych i praktycznego zastosowania wyników detekcji, w tym automatyzacja zasilania baz danych obiektów w obszarze miejskim. Wykorzystanie aspektu przestrzennego ma na celu zdefiniowanie lokalizacji obiektów w terenowym układzie współrzędnych. W tym przypadku głównym problemem badawczym będzie metodyka łączenia wyników z wielu kierunków, eliminacja błędów, a w efekcie zdefiniowanie możliwie jak najdokładniej współrzędnych w układzie terenowym.

Część teoretyczna

1. Metody sztucznej inteligencji w wykrywaniu obiektów

Ludzkie oko bez wątpienia było inspiracją do budowy systemów rozpoznawania obiektów, które aktualnie są stosowane w wielu dziedzinach m.in. medycynie, robotyce czy w pojazdach autonomicznych. Widzenie komputerowe (*ang. Computer Vision, CV*) obejmuje szereg technik mających na celu wydobycie znaczących informacji z danych. Dokładne postrzeganie obrazów, a także rozumienie ich kontekstu przestrzennego i atrybutów jest szeroko wykorzystywane. Pierwsze badania w tym zakresie miały miejsce w latach 50. ubiegłego wieku i dotyczyły testów rozpoznawania podstawowych kształtów. Dalsze postępy technologii w informatyce, rozwój procesorów sprawiły, że obecnie skala wykorzystywania tych systemów jest zdumiewająca.

Dziedzina wykrywania obiektów ewoluowała przez kilka dziesięcioleci z licznymi osiągnięciami. Jednakże jednym z pierwszych algorytmów, który w literaturze (Zou i in., 2023) jest uznawany za przełomowy był algorytm Viola-Jones, wprowadzony w 2001 roku przez Paula Violę i Michaela Jonesa. Algorytm wykorzystywał podejście uczenia maszynowego zwane kaskadami Haara do wykrywania twarzy na obrazach w czasie rzeczywistym. W 2005 roku N. Dalal i B. Triggs zaproponowali deskryptor HOG (*ang. Histogram of Oriented Gradients*) do wykrywania obiektów.

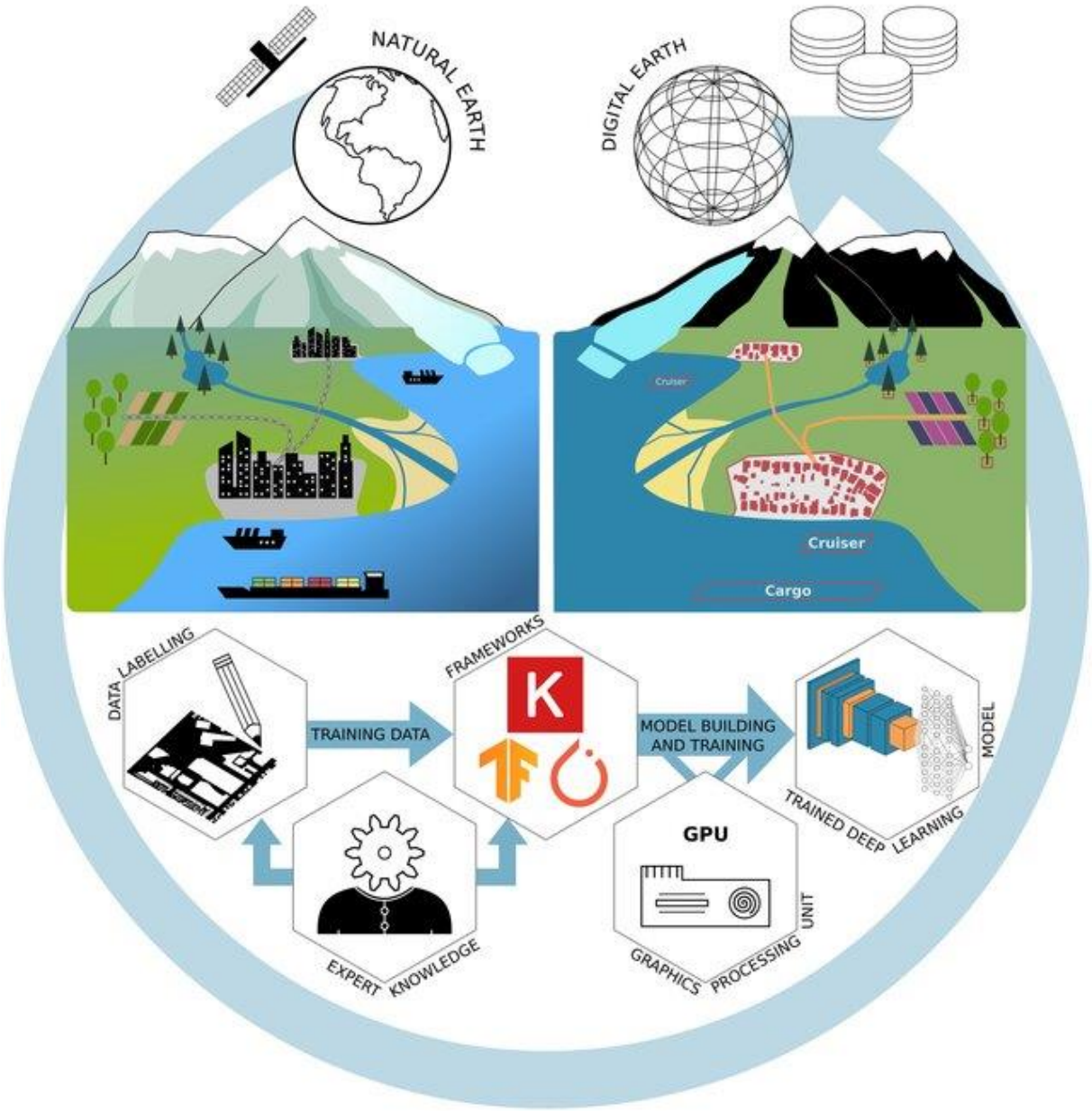
Kolejne lata to rewolucja w widzeniu komputerowym związana z rozwojem algorytmów głębokiego uczenia, umożliwiająca maszynom rozpoznawanie obiektów, śledzenie ruchu i wykonywanie innych złożonych zadań z większą dokładnością niż kiedykolwiek wcześniej. Ważnym osiągnięciem dla zapoczątkowania dominacji metod głębokiego uczenia w CV była pierwsza głęboka sieć konwolucyjna (*ang. convolutional neural network, CNN*) stworzona w 2012 r. przez Alexa Krizhevsky'ego i jego zespół o nazwie AlexNet.

W świetle ewolucji technicznej, która nastąpiła w przeciągu ostatnich 20 lat, głębokie uczenie maszynowe stało się prężnie eksplorowanym przez badaczy obszarem, w szczególności w zadaniach rozpoznawania obiektów, a tematyka wykorzystania metod sztucznej inteligencji wciąż cieszy się dużą popularnością.

1.1. Metody głębokiego uczenia

Bez wątpienia ostatnie lata rozwoju technologicznego przyczyniły się do zrewolucjonizowania sposobu, w jaki maszyny obecnie interpretują świat wizualny. Dzięki wykorzystaniu algorytmów głębokiego uczenia możliwa jest automatyzacja czasochłonnych procesów, co jest szczególnie ważne w dobie generowania ogromnej ilości danych. Przyspieszenie procesów obliczeniowych ułatwia przetwarzanie pozyskanych danych, a w efekcie wydobycie z nich cennych informacji czy nowej wiedzy.

Integracja sztucznej inteligencji z zaawansowanymi zasobami obliczeniowymi zapewnia pełny proces technologiczny w rozpoznawaniu obrazowym, a modele głębokiego uczenia, w szczególności konwolucyjne sieci neuronowe (CNN) zrewolucjonizowały zadania detekcji obiektów i segmentacji obrazu również w dziedzinie obserwacji Ziemi. Poniższy Rysunek 1 przedstawia schemat ukazujący synergię danych geoprzestrzennych i metod DL, dzięki której zapewnione jest działanie mechanizmu od przetwarzania, modelowania, rozpoznawania aż po administrowanie danymi (Qin i Gruen, 2021).



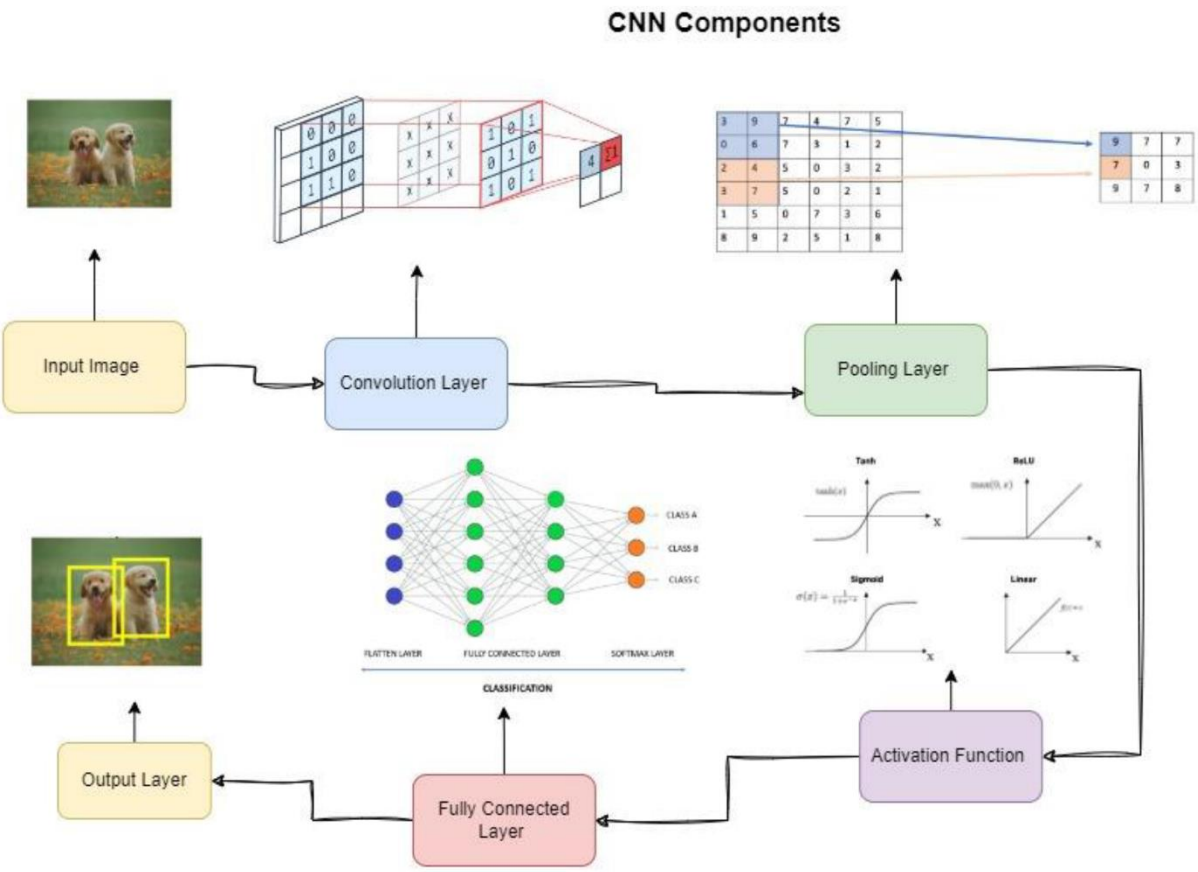
Rysunek 1 Schemat koncepcyjny pozyskiwania baz danych o obiektach na podstawie danych geoprzestrzennych przy użyciu technik głębokiego uczenia. Lewa górna część przedstawia rzeczywisty obraz Ziemi. Prawa górna część przedstawia zdigitalizowaną wersję. Dolna część opisuje proces głębokiego uczenia: tworzenie danych treningowych; budowanie i szkolenie modelu; oraz zastosowanie wyszkolonego modelu do tworzenia baz danych obiektów topograficznych (źródło: Hoeser i in., 2020).

Podstawowym zadaniem modeli głębokiego uczenia jest analiza wprowadzonych danych i wygenerowanie wyniku. Systemy oparte na sztucznej inteligencji są zdolne do rozpoznawania wzorców w danych i automatycznego ich przyswajania. Aby model był niezawodny, dokładny i wydajny konieczne jest przeszkolenie modelu na wysokiej jakości, dużym zbiorze danych.

Niezależnie od zadania, do którego przeznaczony jest model DL, podstawą od której należy zacząć wprowadzenie teoretyczne jest opis klasycznej sztucznej sieci neuronowej (ANN). W skrócie jest to sekwencja w pełni połączonych warstw, od wejściowych przez ukryte

do wyjściowych warstw sztucznych neuronów. Wszystkie neurony w jednej warstwie są połączone z każdym neuronem poprzedniej za pomocą operacji zwanych wagami lub parametrami. W taki sposób poprzez połączenia sieć przenosi wartości z warstwy wejściowej poprzez ukryte warstwy aż do wyjściowej. Aby sieć neuronowa została wytrenowana, należy wprowadzić na wejściu zestaw danych szkoleniowych z etykietami, na podstawie których model zrozumie wzorce i będzie w stanie wykorzystać tę wiedzę, aby wnioskować o nieznanych danych.

W przeciwieństwie do klasycznych modeli ANN, gdzie funkcjami są liniowe połączenia, sieci konwolucyjne (CNNs) pozwalają na uczenie cech danych w bardziej abstrakcyjnym stopniu, uwzględniając lokalne rozmieszczenie dzięki operacjom konwolucyjnym (splot, filtrowanie, *ang. convolution*). W ogólnym stopniu architektura sieci CNN składa się z modułu wejściowego, szkieletu konwolucyjnego, który jest ekstraktorem cech oraz w pełni połączonej warstwy wyjściowej. Rysunek 2 jest przykładem architektury CNN do klasyfikacji obrazów. Modułowa struktura szkieletu, jak i całej architektury sprawia, że sieci CNN są wysoce adaptowalne do różnych zadań.



Rysunek 2 Komponenty konwolucyjnej sieci neuronowej na przykładzie prostego wykrycia obiektów na obrazie. (Źródło: Taye, 2023).

W trakcie ewolucji i opracowywania nowych modeli głębokiego uczenia nastąpił duży rozwój pod względem wydajności i dokładności, co przyczyniło się do wdrożenia w wielu praktycznych zastosowaniach. Analizując literaturę widać, że większość rozwiązań w dziedzinie widzenia komputerowego i głębokiego uczenia dotyczy obrazów dwuwymiarowych. Jednakże potrzeba reprezentowania świata 3D, szczególnie w obszarach miejskich sprawiła, że metody DL zaczęły być narzędziem wykorzystywanym zarówno w przetwarzaniu obrazowym, ale także adaptowanym do danych 3D. Oczywiście modele CNN wykorzystywane do danych 2D i danych 3D (chmur punktów czy modeli 3D) różnią się, jednak podstawowe zasady działania są bardzo podobne i koncentrują się na ekstrakcji cech i hierarchicznym uczeniu się wzorców. Przegląd danych fotogrametrycznych, które wykorzystywane są w głębokim uczeniu wraz z przykładami został opisany w rozdziale 1.2.

1.1.1. Zrozumienie detekcji, segmentacji i klasyfikacji obiektów w widzeniu maszynowym

Zagadnienie wykrywania obiektów niezależnie czy jest realizowane przez bardziej tradycyjne metody uczenia maszynowego, czy głębokie konwolucyjne sieci neuronowe można podzielić w zależności od celu odpowiadającemu właściwemu podejściu. Wyróżnić można między innymi: detekcję, segmentację, klasyfikację, śledzenie i zliczanie obiektów.

Detekcja obiektów to technologia w widzeniu komputerowym, która identyfikuje i lokalizuje obiekty na obrazach cyfrowych. Dzięki wykrywaniu obiektów można określić zarówno co znajduje się na danym obrazie, a także zdefiniować położenie w układzie współrzędnych pikselowych. Zadania, które mogą być realizowane przez detektory są bardzo szerokie od identyfikacji prostych obiektów typu kot czy pies, poprzez wyszukiwanie anomalii na obrazach medycznych, czy analizę wideo w autonomicznych pojazdach.

Pole, w którym model wykrył obiekt nazywane jest ramką ograniczającą (*ang. bounding box*). Prostokąt reprezentujący obiekt daje informacje o lokalizacji obiektu w układzie zdjęcia, ale nie pokazuje dokładnego kształtu obiektu. W przypadku, kiedy niezbędna jest wiedza o dokładnej granicy obiektu należy wykorzystać metody segmentacji obrazu. Segmentacja obrazu idzie dalej niż detekcja i pozwala w szczegółowy sposób określić kontur obiektu, zgodny z dokładnym jego kształtem. Można powiedzieć więc, że segmentacja jest zatem

bardziej szczegółową techniką niż detekcja i pozwala oprócz położenia obiektu poznać również jego kształt oraz dokładny rozmiar. W zadaniu segmentacji część obrazu określająca dany obiekt nazywana jest maską segmentacyjną.

Oprócz detekcji i segmentacji modele mogą również realizować zadanie klasyfikacji. Klasyfikacja obrazu polega na kategoryzacji całego obrazu do jednej klasy. Klasyfikacja zapewnia zrozumienie zawartości obrazu bez możliwości lokalizacji obiektu. Jednakże może służyć jako podstawa dla bardziej złożonych zastosowań, pomagając przykładowo we wstępnej filtracji danych.

Wykrywanie obiektów, segmentacja obrazu i klasyfikacja zapewniają kompleksowy zestaw narzędzi do interpretowania i rozumienia obrazów cyfrowych, dzięki unikalnym możliwościom każdej z metod. Integracja i uzupełnianie się wzajemne jest już zauważalne w rzeczywistych zastosowaniach. Przykładowo, w systemach autonomicznej jazdy detekcja obiektów jest wykorzystywana do identyfikacji innych pojazdów, pieszych czy przeszkód, podczas gdy segmentacja pomaga w precyzyjnym wyznaczaniu granic wykrytych obiektów. Z kolei klasyfikacja może być wykorzystywana do interpretacji sygnalizacji drogowej czy znaków, umożliwiając autonomicznemu pojazdowi podejmowanie decyzji w czasie rzeczywistym.

Dziedzina widzenia komputerowego nadal ewoluuje, a co za tym idzie na pewno dalsze badania będą poczynione przez badaczy i praktyków, aby ulepszać systemy wykrywania obiektów. Kierunek, w którym będzie dążył rozwój wiąże się ze zwiększaniem wydajności, dokładności i użyteczności. Natomiast synergia między wspomnianymi metodami detekcji, klasyfikacji i segmentacji będzie odgrywać kluczową rolę w kształtowaniu przyszłości widzenia komputerowego i jego zastosowań w różnych domenach.

1.1.2. Rola modeli głębokiego uczenia w wykrywaniu obiektów, ze szczególnym uwzględnieniem sieci neuronowych, takich jak YOLO i SAM

Wykorzystanie metod głębokiego uczenia z wizji komputerowej do obserwacji Ziemi jest przedmiotem wielu badań (Hoeser i in., 2020). Większość z nich koncentruje się na rozwoju metod, w których modyfikuje się architekturę modelu, funkcje kosztów i optymalizuje proces szkolenia. Jednak ostatnio rośnie udział badań z zakresu geoinformacji, które stosują metody głębokiego uczenia, aby odpowiedzieć na konkretne pytania badawcze.

Najnowsze kierunki rozwoju w DL to ulepszanie algorytmów pod względem wydajności oraz dostosowywanie struktury modeli, a także wykorzystywanie modeli wyuczonych na dużych zbiorach danych jako podstawy do ewentualnego dotrenowania. W świetle wyzwań związanych z wykrywaniem obiektów, w szczególności gdy mowa o wysokorozdzielczych zobrazowaniach lotniczych i zastosowaniach w obserwacji Ziemi duże znaczenie ma zdolność modelu do wykrywania małych obiektów, które zajmują stosunkowo mały obszar na obrazie.

Wiele modeli CNN opracowanych do tej pory (wśród nich modele takie jak: Faster R-CNN, SSD (Single Shot MultiBox Detector), RetinaNet, Spatial Pyramid Pooling Networks (SPP-Net), Region-based CNN (R-CNN), CornerNet, RefineDet, YOLO (ang. You Only Look Once)) osiągnęło dużą skuteczność w wykrywaniu obiektów, segmentacji instancji i zadaniach klasyfikacyjnych. Jednakże do jednych z najczęściej używanych w kontekście detekcji obiektów można zaliczyć modele SSD i YOLO, co może być związane z ich dobrą wydajnością i obszerną dokumentacją oraz szeregiem eksperymentów przeprowadzonych z ich użyciem (Hoeser i in., 2020).

Wang i in., 2023 zoptymalizowali model YOLOv8 do wykrywania obiektów na zdjęciach UAV i wykazali, że uzyskał wyższą dokładność względem innych modeli. Aby zaradzić problemom związanym z wykrywaniem małych obiektów, Wu i Dong (2023) zaproponowali algorytm YOLO-SE, który jest modyfikacją modelu YOLOv8, osiągając dokładność (mAP – *ang. mean Average Precision*) na poziomie 86.5%. Podobne rozwiązanie ulepszenia modelu w wersji 8 przedstawił Yi i in. (2023), nazywając nowy model LAR-YOLOv8, dla którego wartość wskaźnika mAP w porównaniu z innymi przetestowanymi modelami była wyższa o kilka punktów procentowych.

O ile pierwsze wersje modelu YOLO służyły głównie do zadania detekcji, to wraz z rozwojem nastąpiła modyfikacja i kolejne wersje były ulepszone tak, by oprócz detekcji obiektów na obrazie mogły również wykonywać jego segmentację. Jednakże w kontekście zadań segmentacji dominowała sieć U-Net, co potwierdza przegląd (Hoeser i in., 2020), w którym wykazano, że spośród 261 prac odnoszących się do segmentacji obrazów za pomocą głębokiego uczenia w obserwacji Ziemi aż 33% bazowało na tym właśnie modelu.

Przełom w segmentacji obrazu, który nastąpił dzięki wprowadzeniu innowacyjnego modelu SAM (*Segment Anything Model*) zaowocował wdrożeniem tego modelu w obszar bazujący na danych geoprzestrzennych. Potencjał wykorzystania i przyszłe kierunki badań zostały nakreślone przez autorów Osco i in. (2023), zaś Sultan i in. (2023) opracowali GeoSAM, który

został dostosowany do segmentacji infrastruktury drogowej i pieszej na podstawie zobrazowań satelitarnych. Wykorzystanie dwóch nowatorskich koncepcji - SAM-Generated Object (SGO) i SAM-Generated Boundary (SGB) zostało przedstawione w pracy (Ma i in., 2023), gdzie wyniki eksperymentów również wykazały skuteczność proponowanej metody. Widać więc, że prawdopodobnie w najbliższym czasie będzie powstawało coraz więcej prac ukazujących możliwości adaptacji modelu SAM w obszarze geoinformacji i obserwacji Ziemi.

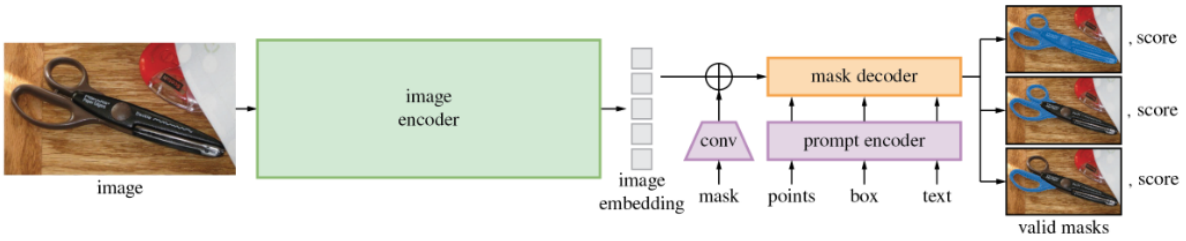
Przywołując najnowsze przykłady z literatury na temat wykorzystania modeli takich jak YOLO i SAM należy również wspomnieć o podejściu, które prawdopodobnie również będzie pogłębiane, a mianowicie integracja modeli i wykorzystywanie potencjału obu algorytmów w jednym procesie. W jednej z prac (Khatua i in., 2024) przedstawiona została metodyka, w której model SAM wykorzystuje do segmentacji obrazu sklasyfikowane dane wyjściowe w postaci ramek ograniczających wygenerowanych przez YOLOV8 jako podpowiedzi. Dzięki takiej synergii możliwe jest osiągnięcie lepszych wyników.

Bazując na najnowszych trendach przybliżonych powyżej zdecydowano, że najnowsze rozwiązania zostaną wdrożone w prace eksperymentalne. Zarówno model YOLO, jak i model SAM zostaną wykorzystane w badaniach. Poniżej natomiast przybliżono charakterystykę i opis architektury obu modeli.

Segment Anything Model (SAM)

W kwietniu 2023 r. firma META wydała wspomniany model o nazwie Segment Anything Model (SAM) (Kirillov i in., 2023). Model ten jest używany jest do segmentacji obrazów i można uznać, że obecnie jest najnowocześniejszym (*ang. state-of-the-art*) modelem segmentacji obrazu. Ponadto przewyższa inne podejścia z tego obszaru, gdyż został przeszkolony na wysokiej jakości zbiorze danych zawierającym 11 milionów obrazów i ponad miliard masek.

SAM składa się z bardzo wymagającego głębokiego kodera obrazu, wykorzystującego zaawansowane maskowane autoenkodery (He i in., 2022) oraz wstępnie wytrenowanego modelu transformatora wizyjnego (Dosovitskiy i in., 2020). SAM składa się z trzech komponentów (Rysunek 3): kodera obrazu (*ang. image encoder*), elastycznego kodera podpowiedzi (*ang. prompt encoder*) i szybkiego dekodera maski (*ang. mask decoder*).



Rysunek 3 Architektura SAM składa się z trzech komponentów, które wzajemnie się uzupełniają w celu zwrócenia prawidłowej maski segmentacji (źródło: Kirillov i in., 2023).

Na najwyższym poziomie koder obrazu generuje jednorazowe osadzenia obrazu (*ang. one-time embedding*) i wyodrębnia jego podstawowe cechy, które służą jako podstawa do późniejszej segmentacji. Koder podpowiedzi konwertuje natomiast podpowiedzi użytkownika na wektory osadzania (*ang. embedding vectors*) w czasie rzeczywistym. Koder interpretuje różne formaty podpowiedzi (*ang. prompts*) i transformuje je na format zrozumiały dla modelu. SAM uwzględnia dwa zestawy podpowiedzi: rzadkie (punkty, pola, tekst) i gęste (maski). Punkty i pola są reprezentowane przez kodowanie pozycyjne i dodawane z wyuczonymi osadzeniami dla każdego typu podpowiedzi. Swobodne podpowiedzi tekstowe są reprezentowane za pomocą gotowego kodera tekstu. Gęste podpowiedzi, takie jak maski, są osadzane za pomocą splotów i sumowane element po elemencie za pomocą osadzania obrazu. Dekoder masek przewiduje maski segmentacji na podstawie osadzeń z koderów obrazu i podpowiedzi. Generowane są dokładne maski, które identyfikują obiekt lub region określony przez użytkownika.

Możliwości uogólniania SAM sprawiają, że jest to rozwiązanie, które można wykorzystywać bez potrzeby dodatkowego szkolenia i konieczności posiadania dobrze oznakowanych danych do trenowania. SAM jest w stanie automatycznie identyfikować i generować maski dla wszystkich obiektów obecnych na obrazie. Przewiduje również maski obiektów na podstawie podpowiedzi wskazujących obiekt docelowy. Inżynieria podpowiedzi (*ang. prompt engineering*) jest to koncepcja związana z metodami AI i jest wykorzystywana często w modelach przetwarzania języka naturalnego (*ang. Natural Language Processing*).

Celem projektu było rozpowszechnienie segmentacji do nowych zadań, dlatego też zespół META udostępnił model SAM na otwartej licencji (Apache 2.0). Ponadto udostępniony został również największy w historii zbiór danych segmentacji (Segment Anything 1-Billion mask dataset (SA-1B)), aby wspierać dalsze badania nad modelami widzenia komputerowego.

Uzupełniając szczegółowe informacje, w modelu SAM wykorzystywany jest koder obrazu oparty na modelu Vision Transformer (VIT) wstępnie wytrenowanym przy użyciu podejścia

Masked Auto Encoders (MAE). W ramach projektu *open source segment-anything* dostępne są wstępnie wytrenowane wagi SAM dla trzech różnych wariantów architektury ViT, które różnią się skalą: ViT-B SAM, ViT-L SAM, ViT-H SAM.

ViT-H znacznie przewyższa ViT-B, gdyż posiada więcej parametrów, ale jednocześnie jest bardzo wymagający obliczeniowo i często używany w zaawansowanych zastosowaniach. Enkodery te mają różną liczbę parametrów: ViT-B ma 91 mln, ViT-L 308 mln, a ViT-H 636 mln parametrów.

Model YOLOv8

Pierwszy raz model YOLO (*You Only Look Once*) został wprowadzony na rynek w roku 2015. Opracowany przez Josepha Redmona i Alego Farhadiego na Uniwersytecie Waszyngtońskim model szybko zyskał na popularności dzięki szybkości i dokładności. Rok później została wydana ulepszona wersja YOLOv2. Przez następne lata sukcesywnie powstawały kolejne iteracje YOLO, wprowadzając stopniowe coraz większe ulepszenia i innowacje funkcje, aż do najnowszej wersji, YOLOv9.

W momencie wykonywania eksperymentów najnowocześniejszym modelem do detekcji obiektów i segmentacji obrazu był stworzony przez Ultralytics[1] model YOLOv8 (wydany w styczniu 2023 r.), dlatego też tę wersję modelu wdrożono do części eksperymentalnej. Model ten posiada kilka rozbudowanych funkcji, o których należy krótko wspomnieć.

Po pierwsze model ten może być wykorzystywany w łatwy sposób z modelami wstępnie wytrenowanymi na dużych zbiorach danych np. COCO. Poza tym może być dostosowywany do różnych wymagań dzięki elastycznej architekturze.

YOLOv8 jest tzw. modelem „*anchor-free*”. Aby lepiej zrozumieć ideę należy wyjaśnić czym są tzw. *anchor boxy*, które są wykorzystywane przez większość tradycyjnych rozwiązań. Jest to wstępnie zdefiniowane pole ograniczające o stałym rozmiarze, które służy jako odniesienie do przewidywania występowania danego obiektu na zdjęciu. Podczas treningu model uczy się przewidywać przesunięcie i zmiany rozmiaru potrzebne do dopasowania ramek ograniczających wokół rzeczywistych obiektów.

Modele takie jak YOLOv8 nie bazują na tych wstępnie zdefiniowanych polach (*anchor boxes)*, przewidują natomiast bezpośrednio punkty określające środek położenia obiektu. Predykcja

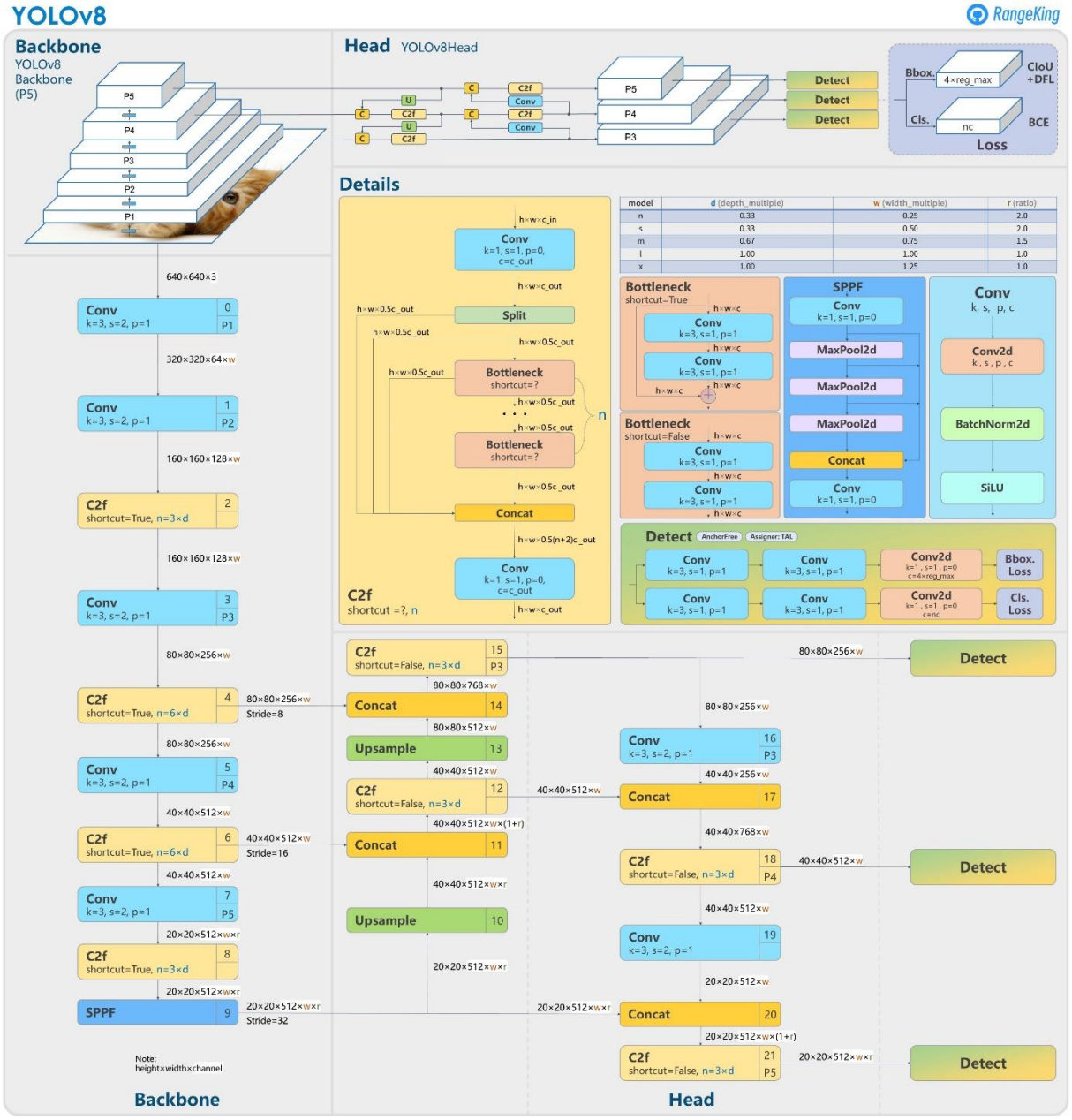
[1] https://github.com/ultralytics/ultralytics

opiera się zatem na rzeczywistych wymiarach obiektu, a nie na korektach wynikających z wcześniej zdefiniowanej ramki. Wydajność, prostota i elastyczność modeli „*anchor-free*" prowadzi do krótszego procesu uczenia i potencjalnie dokładniejszego wykrywania.

Kolejną innowacyjną funkcją wdrożoną w YOLOv8 jest wykorzystanie mozaikowego rozszerzania danych (*ang. mosaic data augmentation*), które łączy cztery obrazy w jeden, aby zapewnić modelowi lepsze informacje kontekstowe.

Architektura modelu została zaktualizowana, zastępując poprzedni moduł C3 nowym modułem C2f. Dzięki nowej strukturze możliwe było zmniejszenie parametrów modelu, rozszerzenie pola recepcyjnego i zwiększenie wydajności obliczeniowej. W YOLOv8 "głowa" modelu, która tradycyjnie zarządzała zarówno klasyfikacją obiektów, jak i przewidywaniem ich lokalizacji, obsługuje teraz te zadania oddzielnie. To rozdzielenie pozwala modelowi stać się bardziej wyspecjalizowanym w każdej funkcji.

Poniższy schemat (Rysunek 4) wykonany przez użytkownika GitHub – RangeKing przedstawia szczegółową wizualizację architektury YOLOv8, która składa się z kilku kluczowych komponentów. Szkielet (*ang. backbone*) to seria warstw konwolucyjnych, które wyodrębniają odpowiednie cechy z obrazu wejściowego. Warstwa SPPF i kolejne warstwy konwolucyjne przetwarzają cechy w różnych skalach, natomiast warstwy Upsample zwiększają rozdzielczość map cech. Moduł C2f łączy cechy wysokiego poziomu z informacjami kontekstowymi w celu poprawy dokładności wykrywania. Moduł detekcji wykorzystuje zestaw warstw konwolucyjnych i liniowych do mapowania wielowymiarowych cech.



Rysunek 4 Struktura modelu detekcji obiektów YOLOv8. Prostokąty reprezentują warstwy, z etykietami opisującymi typ warstwy (Conv, Upsample itp.) i wszelkie istotne parametry (rozmiar jądra, liczba kanałów itp.). Strzałki reprezentują przepływ danych między warstwami, a kierunek strzałki wskazuje przepływ danych z jednej warstwy do drugiej. Źródło: https://github.com/ultralytics/ultralytics/issues/189.

Wykorzystując możliwości takich modeli jak YOLO, SAM czy innych podobnych modeli głębokiego uczenia, możliwe jest zwiększenie wydajności i dokładności wykrywania obiektów. Ponadto zdolność do generalizacji w szerokim zakresie obiektów i obrazów sprawia, że takie rozwiązania są szczególnie atrakcyjne do wdrożenia.

Wybór obu modeli (SAM i YOLOv8) na potrzeby prac badawczych był podyktowany zdolnością adaptacyjną oraz tym, że w przeciwieństwie do tradycyjnych modeli, które zwykle

wymagają dużych ilości danych, te mogą wyuczyć się z niewielkiej liczby przykładów. Co więcej, decyzja była poparta szerokim przeglądem literatury, gdzie oba modele były przywoływane jako najnowocześniejsze podejścia w zakresie wykrywania obiektów.

1.1.3. Metody data augmentation

Istotnym punktem uczenia głębokiego jest konieczność posiadania danych wejściowych, które są zróżnicowane i dobrej jakości, aby model był w stanie osiągnąć wysoką dokładność. Zdarza się jednak czasem, że rozmiar próbki danych szkoleniowych jest zbyt mały. W takim wypadku jednym z rozwiązań jest oczywiście zebranie większej ilości danych, co wydaje się być najbardziej skuteczną metodą, ale równocześnie bardzo czasochłonną, gdyż najczęściej sprowadza się do manualnej pracy. Jednakże istnieją też inne rozwiązania pozwalające na rozszerzenie zbiorów uczących z wykorzystaniem bardziej automatycznych metod.

Jedną z najczęściej stosowanych technik jest rozszerzanie zbiorów uczących z wykorzystaniem metod augmentacji danych (*ang. data augmentation)*. W szczególności dużym zastosowaniem cieszą się te metody w przypadku przetwarzania obrazów.

Rozszerzanie zbioru danych dla zdjęć polega na modyfikacjach obrazów zawartych w zbiorze treningowym. Zasadniczo jest to generowanie prawdziwie wyglądających fałszywych danych. W efekcie tworzone są różne wersje podobnej zawartości danych w celu zwiększenia różnorodności przykładów treningowych dla modelu. Proces ten może znacząco poprawić zdolność modelu do generalizacji, a tym samym działać bardziej efektywnie na wcześniej niewidzianych obrazach. Ponadto uwzględnienie metod augmentacji danych umożliwia generowanie przypadków, które nie wystąpiły w zbiorze uczącym, a mogą wystąpić w świecie rzeczywistym.

Pośród metod sztucznego poszerzania zbiorów uczących w odniesieniu do obrazów wyróżnia się następujące metody:

- flip – odbicie lustrzane wokół osi x lub y,
- rotation – obrót obrazu o określony kąt (zgodnie lub przeciwnie do ruchu wskazówek zegara),
- crop – przycinanie obrazu, czyli w zasadzie ekstrakcja podzbioru obrazu (może być wykorzystywane do poprawy uogólniania modelu),

- shear – zniekształcenie obrazu wzdłuż osi, głównie w celu utworzenia lub skorygowania kątów postrzegania,
- blur – rozmycie, najczęściej wprowadzane jest rozmycie gaussowskie do obrazu,
- exposure – dostosowanie ekspozycji w celu rozjaśnienia lub przyciemnienia obrazu,
- noise – wprowadzenie szumu do obrazu (powszechnie stosowany jest szum „salt and pepper", w którym piksele obrazu są losowo zmieniane na całkowicie czarne lub całkowicie białe),
- saturation - zmiana intensywności kolorów na obrazie.

Jak widać metody data augmentation można podzielić na te oparte na kolorach (np. zmiana kontrastu) i oparte na geometrii (np. obracanie). Oczywiście nie ma określonego zestawu metod, które powinny być zawsze zaimplementowane, dzięki któremu istnieje pewność, że wydajność modelu natychmiast się poprawi. Co więcej, w rzeczywistości zdarza się tak, że niektóre metody mogą mieć negatywny wpływ na model. Przykładowo obracanie obrazów może być nielogiczne w zadaniach zależnych od kolejności np. przy interpretacji rozpoznawanego tekstu. Znajomość kontekstu przygotowywania danych do uczenia i wnioskowania o modelu jest wymagana do podejmowania świadomych decyzji dotyczących wdrażania metod sztucznego rozszerzania zbiorów uczących.

Stosowanym podejściem do pracy z modelem z wykorzystaniem nowych danych najczęściej jest wstępne wytrenowanie modelu bez żadnych metod augmentacji danych. Pozwala to ocenić jakość bazowego zbioru danych. Jeśli do zestawu treningowego na początku zostaną dodane sztucznie wytworzone przykłady trudno będzie ocenić ewentualnie występujące problemy. W takim przypadku bardziej adekwatnym rozwiązaniem będzie wstępna kontrola jakości przygotowanego zestawu uczącego, zbadanie równowagi klas i reprezentacji danych. Gdy pierwsza iteracja trenowania modelu nie wykaże dużych błędów można dodać nowe dane i poszerzyć zbiór, aby poprawić wydajność modelu.

Metody augmentacji danych mogą być stosowane do całego obrazu lub w przypadku detekcji obiektów, do ramek ograniczających obiekt zainteresowania. W przypadku zastosowania tzw. *bounding box augmentation* tworzone są nowe dane treningowe, ale zmieniana jest jedynie zawartość ramki ograniczającej obiekt. W taki sposób możliwa jest przykładowo zmiana jasności obiektu względem tła. Ciekawym przykładem jest rozmycie w stosunku do tła, które może mieć zastosowanie do wykrywania poruszających się obiektów.

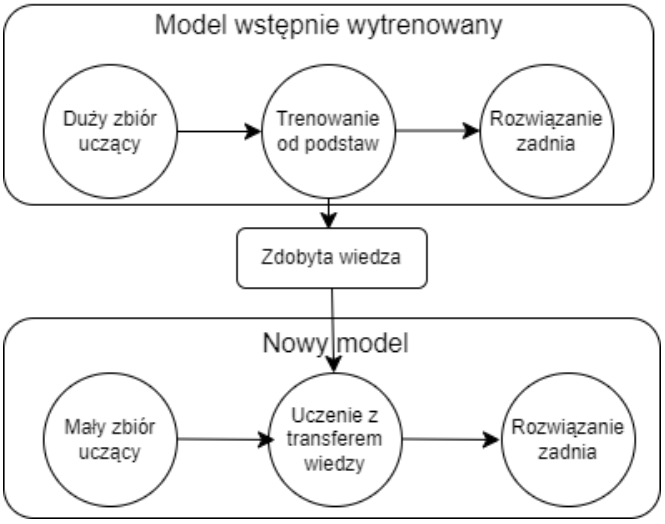
W artykule (Zoph i in., 2020) naukowcy zbadali rozszerzanie zbioru treningowego w oparciu o metody *data augmentation* zastosowane do ramek ograniczających. W pracy zostało wykazane, że modyfikacje oparte wyłącznie na ramkach ograniczających skutkują systematyczną poprawą, szczególnie w przypadku modeli dopasowanych do małych zbiorów danych.

Rysunek 5 Przykłady augmentacji danych. A) odbicie lustrzane, B) ekspozycja 10%.

1.1.4. Transfer Learning

W stale rozwijającej się dziedzinie AI potrzeba skutecznych metod trenowania wydajnych modeli jest obecna, a jednym z najbardziej czasochłonnych procesów opracowania, jak zostało wcześniej nakreślone jest ręczna adnotacja danych treningowych. Aby sprostać tym wyzwaniom naukowcy opracowują coraz to nowsze techniki dzięki którym można znacząco przyspieszać zarówno procesy tworzenia zbiorów uczących, jak i sam przebieg uczenia modelu.

Jednym z takich podejść, które aktualnie jest bardzo często w wielu projektach wykorzystywane jest tzw. transfer wiedzy (*ang. transfer learning*). Jest to technika, w której nowy model uczy się rozwiązać jakieś zadanie na podstawie wiedzy zdobytej przez inny model (Rysunek 6). W praktyce wygląda to tak, że wykorzystuje się wstępnie wytrenowane modele jako podstawę nowego modelu, który można następnie dostroić z wykorzystaniem mniejszego zbioru danych.



Rysunek 6 Schemat działania metody transfer learning.

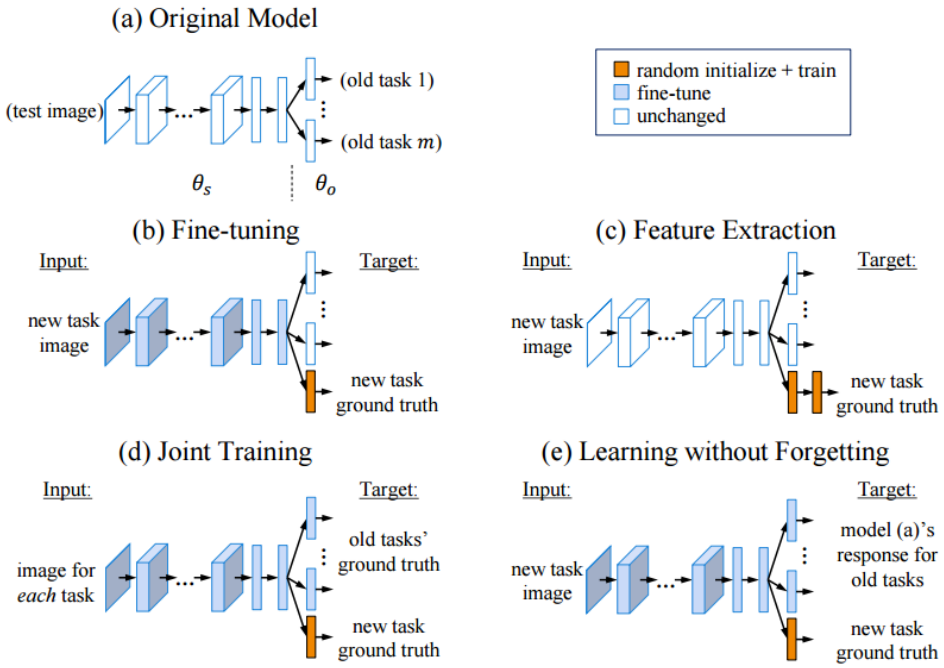
Konwolucyjne sieci neuronowe, które są najczęściej stosowane do zadań detekcji obiektów wyodrębniają cechy z obrazów zaczynając od rozpoznawania ogólnych cech. Początkowe warstwy modelu uczą się więc właściwości, które mogą być wspólne w niektórych domenach. Końcowe warstwy modelu są dostosowywane do rozróżniania cech istotnych w konkretnym zadaniu.

W odniesieniu do tematyki wykrywania obiektów proces ten zazwyczaj rozpoczyna się od wykorzystania modelu bazowego, który został wstępnie wytrenowany na dużym, ogólnym zbiorze danych. Przykładem takiego zbioru jest ImageNet. Taki wstępnie wytrenowany model posiada już zapisane wzorce pozwalające zrozumieć podstawowe cechy, na przykład kształty

czy krawędzie. Gdy model jest wstępnie wytrenowany, stosując go do innego zadania zmieniane są jedynie końcowe warstwy modelu, które muszą zostać ponownie wytrenowane na podstawie nowych danych lub dostrojone za pomocą takiego zestawu.

W obszarze związanym z detekcją obiektów z wykorzystaniem konwolucyjnych sieci neuronowych wyróżnia się różne metody (Rysunek 7), a dwa główne scenariusze zastosowania transfer learningu obejmują:

- dostrojenie wstępnie wytrenowanego modelu (*ang. fine-tuning*) – polega na modyfikacji wstępnie wytrenowanego modelu, dostrajanie może być wykonane na całej sieci lub ograniczone do niektórych warstw, podczas treningu na nowych danych w tej strategii wszystkie wagi warstw wstępnie wytrenowanego modelu są aktualizowane, z wyjątkiem wag ostatnich warstw dla oryginalnego zadania;
- wykorzystanie wstępnie wytrenowanego modelu jako ekstraktora cech (*ang. feature-extraction*) – polega na tym, że ze wstępnie wytrenowanego modelu wykorzystywane są tylko początkowe warstwy do ekstrakcji przydatnych cech; w tej strategii tylko wagi nowo dodanych ostatnich warstw zmieniane są podczas uczenia.



Rysunek 7 Zasady działania kilku różnych podejść transfer learning: a) model wyuczony na podstawie oryginalnych danych; b) dostrajanie, które polega na tym, że model wyuczony wcześniej jest dostosowywany do nowego zadania poprzez aktualizację niektórych parametrów; c) ekstrakcja cech polega na wykorzystaniu wstępnie wytrenowanego modelu jako ekstraktora cech, a trening przebiega tylko z wykorzystaniem ostatnich warstw; d) metoda ta obejmuje jednoczesne szkolenie w zakresie nowego zadania, jak i źródłowego; e) uczenie się bez zapominania, którego celem jest nauczenie sieci, która może dobrze radzić sobie zarówno ze starymi zadaniami, jak i nowymi, gdy dostępne są tylko dane dotyczące nowego zadania bez konieczności ponownego trenowania (źródło: Li i Hoiem, 2017).

Ważną kwestią jest jednak to, że zadania realizowane przez model wstępnie wytrenowany i ten nowy douczany powinny być w miarę zbliżone dziedzinowo. Jeśli dane, na podstawie których model został wstępnie wytrenowany znacznie różnią się od nowego zbioru danych, uczenie transferowe może nie zadziałać. Oczywiście może zdarzyć się tak, że mimo pokrewnego zadania zaimplementowanie uczenia przez transfer nie przyniesie oczekiwanych efektów. Taki scenariusz jest prawdopodobny przy zagadnieniach wymagających bardzo dużych zbiorów danych. W takim wypadku należy rozważyć czy uczenie od podstaw nie byłoby bardziej rozsądnym rozwiązaniem.

Stosowanie transferu wiedzy pozwala przeszkolić nowy model do konkretnego celu nie wymagając przy tym dużej ilości danych uczących. Czas wymagany na wyszkolenie modelu zmniejsza się przy zastosowaniu metod transfer learning, gdyż wstępnie wytrenowany model już raz nauczył się znacznej ilości informacji, więc ilość danych potrzebna do dotrenowania modelu jest mniejsza niż przy trenowaniu od podstaw. Uczenie przez transfer wiedzy pozwala na osiągnięcie lepszych wyników przy znacznie krótszym czasie uczenia oraz oszczędzania zasobów obliczeniowych. Zaletą jaka z tego wynika jest również dostępność dla użytkowników posiadających ograniczone zasoby. Rozwój tej techniki zdecydowanie miał wpływ na przyspieszenie udoskonalania i poprawę wydajności modeli w zakresie wykrywania obiektów w różnych branżach.

W odniesieniu do prac badawczych, wspomniane metody głębokiego uczenia – model SAM i YOLOv8 zostaną wykorzystane w eksperymentach. Pierwsze podejścia będą skupiały się bezpośrednio na zastosowaniu wstępnie wytrenowanych modeli do nowego zadania detekcji bez dostrajania. W kolejnych etapach podjęta zostanie próba dostrojenia modelu i przetrenowanie na nowym zbiorze danych. Opisane powyżej metody sztucznego poszerzania zbiorów uczących oraz uczenie transferowe zostaną wykorzystane na potrzeby badań. Zastosowanie takiego podejścia jest możliwe dzięki dużej dostępności wstępnie wytrenowanych modeli i narzędziom takim, jak Roboflow[2], które ułatwiają proces przygotowania zbiorów do treningu wraz z rozszerzeniem danych metodami data augmentation.

[2] https://roboflow.com/

1.2. Dane fotogrametryczne w dziedzinie głębokiego uczenia

Automatyczna detekcja, segmentacja czy klasyfikacja ogromnej ilości pozyskanych danych stała się ważnym zadaniem w ostatniej dekadzie. Największy postęp w zakresie rozpoznawania obrazów nastąpił w przypadku scen naturalnych (w rozumieniu zdjęć naziemnych), czy też filmów. Jednakże tego typu dane często nie posiadają odniesienia przestrzennego. Oczywiście nie każde zastosowanie wymaga informacji przestrzennej, jednak potencjał jaki niesie za sobą wykorzystywanie wieloźródłowych danych fotogrametrycznych i teledetekcyjnych jest ogromny, szczególnie w zastosowaniach mapowania miejskiego (Laupheimer i Haala, 2021).

Zgodnie z definicją, w fotogrametrii badane są kształty, lokalizacje, rozmiary i wzajemne relacje rzeczywistych obiektów na podstawie danych, a więc oprócz geometrii może być brana pod uwagę również wartość semantyczna. Podstawowymi reprezentacjami danych są obrazy cyfrowe i chmury punktów LiDAR (*ang. Light Detection and Ranging*) lub wygenerowane z gęstego dopasowania obrazów (*ang. dense image matching, DIM*), które pozyskuje się z pułapu lotniczego, satelitarnego oraz z dronów. Dane mogą być następnie przetwarzane do innych użytecznych produktów takich jak numeryczne modele terenu (NMT), numeryczne modele pokrycia terenu (NMPT), modele siatkowe 3D (*ang. mesh models*), modele bryłowe budynków, mapy pokrycia terenu czy ortofotomapy. Widać więc, że w dziedzinie głębokiego uczenia fotogrametria zapewnia wysokiej jakości, wartościowe, ustrukturyzowane zbiory danych, na podstawie których trenowane są modele.

Przybliżając tematykę wykorzystania danych fotogrametrycznych w dziedzinie głębokiego uczenia należy wspomnieć o uczeniu maszynowym (*ang. Machine Learning, ML*), które również odgrywa ważną rolę w problemach rozpoznawania obiektów, czy klasyfikacji obrazów w fotogrametrii i teledetekcji. ML obejmuje m. in. metody takie jak: regresja logistyczna (Menard, 2018), maszyna wektorów nośnych (*ang. Support Vector Machine, SVM*) (Wang, 2005), drzewa decyzyjne (Friedl i Brodley, 1997). Wspomniane metody stosowano na przykład do celów klasyfikacja pokrycia terenu i użytkowania gruntów z wykorzystaniem obrazów satelitarnych (Ma i in., 2017).

Jednakże, jak już wcześniej zostało podkreślone, to metody głębokiego uczenia, w szczególności konwolucyjne sieci neuronowe (CNNs) są najnowocześniejszymi metodami segmentacji, detekcji i klasyfikacji (Song i in., 2019; Minaee i in., 2021; Zou i in., 2023). Ponadto wyniki wielu badań naukowych z tego obszaru pokazują, że rozwiązania wykorzystujące CNNs przewyższą inne klasyczne klasyfikatory pod względem dokładności.

Poniższy opis skupia się na przeglądzie danych fotogrametrycznych wykorzystywanych w dziedzinie głębokiego uczenia.

1.2.1. Obrazy 2D

Tradycyjnie, głębokie uczenie zostało opracowane do analizy obrazów RGB. Najczęściej stosowanymi są obrazy optyczne o wysokiej rozdzielczości przestrzennej. Jednakże udział badań wykorzystujących zobrazowania wielospektralne, radarowe czy termalne zyskuje na popularności (Hoeser i in., 2020). Ponadto duża liczba aplikacji wykorzystuje sceny satelitarne, gdyż teledetekcja zapewnia dane o znacząco wysokiej (np. WorldView, GeoEye itp.) rozdzielczości, która w przypadku niektórych zastosowań jest wystarczająca. Zaletą zdjęć satelitarnych jest również to, że są darmowym źródłem danych (np. Landsat czy Sentinel).

Obok scen satelitarnych oczywiście szeroko stosowanymi są zdjęcia lotnicze i dane obrazowe UAV, zwłaszcza w mniejszych regionach czy środowisku miejskim (Schlosser i in., 2020). Rozwiązania wykorzystujące zdjęcia lotnicze i UAV są przeznaczone zarówno do zadań segmentacji semantycznej, klasyfikacji, jak i detekcji obiektów.

Wiele z publikowanych metod skupia się również na klasyfikacji produktów typu ortofotomapa czy true-orto powstałych z gęstego dopasowania zdjęć (Schlosser i in., 2020). Pozwala to na tworzenie baz danych 2D, jednakże tego typu dane mają swoje ograniczenia związane ze zniekształceniami.

Zastosowanie konwolucyjnych sieci neuronowych do klasyfikacji, segmentacji i detekcji zdjęć lotniczych pionowych jest tematem dość szeroko przebadanym, o czym świadczy chociażby liczba benchmarków powstałych na przełomie ostatnich lat (Zhao i in., 2021; Ding i in., 2021; Hoeser i Kuenzer, 2020).

Znacznie mniej aplikacji jest z zakresu zdjęć ukośnych, szczególnie w przypadku obrazów lotniczych. Świadczy o tym liczba ogólnodostępnych zbiorów danych. Tematyka ta została bardziej szczegółowo opisana w rozdziale 2.3.1.

1.2.2. Chmury punktów 3D

Podczas gdy dane 2D zawierają bogate informacje tematyczne, istnieją ograniczenia jeśli chodzi o wykorzystanie tego rodzaju danych w przypadku pozyskiwania szczegółowych informacji geometrycznych. Naprzeciw tym wyzwaniom wychodzą gęste trójwymiarowe chmury punktów, które mogą służyć do automatycznej identyfikacja kształtów i segmentacji obiektów.

Obszarem, w którym głębokie uczenie z wykorzystaniem chmur punktów jest implementowane na szeroką skalę są aplikacje podejmujące decyzje w czasie rzeczywistym, takie jak autonomiczna jazda i nawigacja w środowisku miejskim.

Mnogość pozyskiwanych danych 3D oraz zdolność sensorów do pozyskiwania chmur punktów o wysokiej gęstości zachęciła badaczy do rozwiązywania zadań związanych z rozumieniem scen miejskich z wykorzystywaniem tego typu danych, co było podyktowane potrzebą gromadzenia informacji 3D o obiektach w świecie rzeczywistym.

Zwiększone zainteresowanie danymi 3D, a także postępy w opracowywaniu algorytmów DL do pracy na takich zbiorach danych, stworzyły zapotrzebowanie na dostęp do dobrze oznakowanych zbiorów treningowych i testowych (Zolanvari i in., 2019). Dlatego też podobnie, jak w przypadku obrazów 2D powstaje coraz to więcej zbiorów danych do segmentacji i klasyfikacji chmur punktów 3D, pozwalających na niezależne porównywanie wyników badań (Zachar i in., 2023).

1.2.3. Modele siatkowe i bryłowe 3D

Oprócz chmur punktów istnieją również inne formy danych 3D, takie jak modele siatkowe (*ang. mesh model*) i modele bryłowe zgodne ze standardem CityGML. Oteksturowane modele siatkowe są aktualnie tworzone głównie na podstawie gęstego dopasowania ukośnych zdjęć lotniczych i stają się coraz bardziej popularne jako metoda reprezentacji środowiska miejskiego. Uzupełnianie modeli mesh o dodatkowe informacje jest najczęściej realizowane przez algorytmy AI. Chociaż uzyskanie dobrych wyników segmentacji semantycznej dla modelu miasta mesh 3D jest trudnym zadaniem, to powstają rozwiązania takie jak MeshNet-SP (Zhang i Zhang, 2023), które osiągają dokładności powyżej 90%. Zaś jednym z najczęściej wykorzystywanych benchmarków jeśli chodzi o segmentację semantyczną modeli miast jest zbiór SUM (Gao i in., 2021) pokrywający obszar 4km2 w Helsinkach i składający się z sześciu

klas pokrycia, dla którego w innych przykładowych pracach osiągnięte zostały dokładności na poziomie 91.4% (Wilk i in., 2022) czy 97.1% (Grzeczkowicz i Vallet, 2023) dla różnych podejść wykorzystujących sieci neuronowe.

Rozwój technologiczny kamer oraz sensorów LiDAR rozszerzył potencjał wykorzystania danych fotogrametrycznych oraz produktów, które są wynikiem ich przetwarzania z zaawansowanymi technikami głębokiego uczenia. Ponadto modele DL mogą analizować zmiany w czasie. W tego typu zadaniach wieloczasowe dane fotogrametryczne czy archiwalne zbiory stanowią dobre źródło.

Omawiając różne przykłady zastosowania danych fotogrametrycznych w dziedzinie głębokiego uczenia należy zwrócić uwagę na kilka wyzwań. Po pierwsze dane tego typu wymagają znacznej mocy obliczeniowej. Wynikowa dokładność modeli zależy w dużej mierze od dokładności i rozdzielczości danych. Jednakże potencjał integracji danych fotogrametrycznych i teledetekcyjnych z technikami AI, w szczególności metodami głębokiego uczenia jest widoczny, a wraz ze wzrostem dokładności, zainteresowanie tego typu danymi również będzie rosło.

2. Metody uczenia maszynowego w kontekście przestrzeni miejskiej

Dynamiczny rozwój miast i wysoki stopień wzrostu obszarów zurbanizowanych wymusił intensyfikację ścisłej kontroli i monitorowania. Skutecznymi technikami, które zapewniają szczegółowe dane jest fotogrametria lotnicza, teledetekcja i obrazowanie z drona. Zarówno na podstawie zdjęć, jak i chmur punktów można uzyskać informację o przestrzeni miejskiej i poszczególnych obiektach.

Większość danych, które są wykorzystywane i przetwarzane w mieście posiada odniesienie przestrzenne, dzięki czemu zapewniona jest integracja między różnymi źródłami. Aspekt przestrzenny odgrywa więc kluczową rolę w zarządzaniu miastem i wzmacnia potencjał analityczny. Inwentaryzacja obiektów miejskich i przechowywanie informacji o nich w bazach danych umożliwia monitoring przestrzeni miejskiej oraz wsparcie w podejmowaniu decyzji.

Obiekty miejskie będące przedmiotem zainteresowania obejmują zabudowę, sieć komunikacyjną, tereny zielone, obiekty małej architektury i inne ważne elementy środowiska miejskiego. Wraz z poprawą możliwości sprzętowych zespoły badawcze zostały zachęcone do podejmowania prób opracowania metod automatycznego wykrywania tego typu obiektów na danych pozyskanych przez różne sensory.

Ponadto dzięki temu, że techniki głębokiego uczenia osiągnęły zdumiewający poziom wydajności w wykrywaniu obiektów na obszarach gęstej zabudowy miejskiej z wykorzystaniem danych fotogrametrycznych można powiedzieć, że zastosowanie tego typu rozwiązań będzie wciąż rosło i odpowiadało na kolejne wyzwania, w szczególności w kontekście inteligentnych miast przyszłości.

Oczywiście w ostatnich latach w większości rozwiniętych krajów powstały bazy danych topograficznych 2D. Jednakże wymagają one ciągłej aktualizacji. Stąd można dostrzec potrzebę opracowania narzędzi automatyzacji procesu, które będą stawiać czoła złożoności i różnorodności.

2.1. Zastosowanie uczenia maszynowego do analizy przestrzeni miejskiej

Potwierdzeniem tego, że metody uczenia maszynowego bazujące na danych fotogrametrycznych i teledetekcyjnych są niezbędnym elementem w kontekście analizy przestrzeni miejskiej jest szeroki zakres opracowanych rozwiązań. Biorąc pod uwagę ogromną

różnorodność zastosowań metod AI, poniższy przegląd nie zawiera wszystkich aplikacji, a wybrane które ukazują potencjał wykorzystania metod głębokiego uczenia maszynowego z wykorzystaniem danych fotogrametrycznych z ostatnich lat.

Analizując elementy środowiska miejskiego, można powiedzieć, że budynki są jednymi z najważniejszych obiektów. W kontekście planowania przestrzennego, czy mapowania topografii niezwykle ważne jest uzyskanie dokładnej lokalizacji i kształtów budynków. Wyznaczanie konturów z wykorzystaniem segmentacji semantycznej opartej o CNNs zostało opisane w pracy Li i in. (2019), gdzie wykorzystano sceny satelitarne WorldView-3. Zarówno dane satelitarne, jak i zdjęcia z drona (o rozdzielczości od 2 do 30 cm) zostały wykorzystane do tego samego zadania w pracy Zhao i in. (2021). Eksperymenty wykonane w pracy Farajzadeh i in. (2023) na podstawie ortofotomapy utworzonej z danych UAV i znormalizowanego modelu pokrycia terenu (zNMPT) wykazały możliwości wykorzystania architektury U-Net w celu wyodrębnienia obrysów budynków.

Oprócz informacji o obrysie budynków istotne są informacje o fasadach budynków. W celu wydzielenia poszczególnych elementów fasad budynków w większości aplikacji stosuje się segmentację semantyczną zdjęć ukośnych. Przykładowo sieci neuronowe FC-DenseNet i DeepLabV3+ zostały przetestowane do segmentacji budynków na podstawie zdjęć lotniczych przez Huang i in. (2019). Jednakże w kontekście zadania segmentacji fasad ważną rolę odgrywają dane UAV, które można skutecznie wykorzystywać zarówno w samej analizie elewacji, jak i kontroli uszkodzeń. Zhuo i in. (2019) zaproponowali segmentację fasad, w tym ścian, okien i dachów, na podstawie ukośnych obrazów UAV, jako podstawę do automatycznego generowania modeli 3D. Nowy model YOLOM został natomiast zaproponowany przez Cao (2023) do inspekcji budynków i wykrywania uszkodzeń z wykorzystaniem danych z drona.

Wiele propozycji opartych na DL z danymi lotniczymi 2D zostało przedstawionych w literaturze do identyfikacji pieszych (Tahir i in., 2024), detekcji samochodów (Zachar i in., 2022) czy wykrywania poruszających się obiektów (Heo i in., 2020).

Segmentacja semantyczna pojedynczych gatunków drzew z wykorzystaniem konwolucyjnych sieci neuronowych została zaprezentowana w pracy Lobo Torres i in. (2020). Wykorzystanie obrazów o wysokiej rozdzielczości przestrzennej w celu lokalizacji i klasyfikacji gatunków drzew do uczenia CNN przedstawiono również w publikacji Carnegie i in. (2023), gdzie uzyskana dokładność wyniosła powyżej 90%.

Szczególnie szeroki zakres zastosowań w inteligentnych miastach znalazły modele 3D miast, tworzone głównie za pomocą gęstego dopasowania ukośnych zdjęć lotniczych (Haala i in., 2015). Ponadto coraz częściej wykorzystywane są modele siatkowe z informacją semantyczną, o czym świadczy duża liczba badań, które były przeprowadzone w tym obszarze.

Praca przeglądowa autorstwa Adam i in. (2023) zawiera kompleksowy przegląd najnowszych postępów w wykorzystaniu technik głębokiego uczenia się do semantycznej segmentacji modeli mesh 3D w skali miejskiej. Wilk i in. (2022) w swojej pracy porównali dwa różne podejścia do segmentacji semantycznej modeli mesh 3D wykorzystując przenoszenie klasyfikacji z ukośnych zdjęć lotniczych i chmur punktów na model siatkowy.

Analiza 3D przestrzeni miejskiej to nie tylko stosowanie modeli siatkowych, ale również chmur punktów, pozyskanych z lotniczego skanowania laserowego (ALS), gęstego dopasowania obrazów (DIM), naziemnego skanowania laserowego (TLS) czy też mobilnego skanowania laserowego (MLS). Na podstawie analizy literatury można również stwierdzić, że zainteresowanie danymi 3D jest duże, o czym świadczy wzrost liczby benchmarków do klasyfikacji i segmentacji chmur punktów 3D (Zachar i in., 2023; Wang i in., 2023).

Widać więc, że głębokie uczenie osiągnęło ogromny sukces w klasyfikacji obrazów, wykrywania obiektów i segmentacji semantycznej w odniesieniu do przestrzeni miejskiej. Natomiast aktualne trendy rozwoju to coraz szersze wykorzystanie danych 3D, w tym modeli siatkowych i chmur punktów pozyskanych różnymi technikami.

2.2. Obecny stan wiedzy i najnowsze metody detekcji obiektów architektury miejskiej

Przytoczone wcześniej przykłady nakreśliły jaki potencjał niesie za sobą integracja metod AI, głównie głębokiego uczenia z danymi fotogrametrycznymi i teledetekcyjnymi w rozpoznawaniu obrazowym i wykrywaniu obiektów w dziedzinie obserwacji Ziemi.

Na szczególną uwagę w przestrzeni miejskiej zasługują zdjęcia ukośne, które wraz z obrazami nadirowymi tworzą solidne źródło danych do modelowania 3D obszarów miejskich. Niezaprzeczalne zalety analizy miejskiej z wykorzystaniem zdjęć ukośnych obejmują możliwość mapowania fasad budynków oraz odwzorowania obiektów małej architektury miejskiej. Poniższy przegląd obejmuje najnowsze przykładowe metody detekcji i segmentacji w obrębie obszarów zurbanizowanych ze szczególnym uwzględnieniem segmentacji okien oraz detekcji latarni ulicznych, które będą przedmiotem badań w części eksperymentalnej.

Segmentacja semantyczna budynku ma na celu przypisanie każdemu pikselowi etykiety semantycznej, takiej jak okno, balkon czy drzwi. Jak zostało wspomniane wcześniej, obecnie metody głębokiego uczenia mają znaczącą przewagę jeśli chodzi o dokładności względem tradycyjnych metod takich jak SVM (*ang. Support Vector Machine)* czy RF (*ang. Random Forest*), dlatego poniższy przegląd uwzględnia rozwiązania oparte o DL.

Huang, S. (2019) w swojej pracy użył głębokich sieci neuronowych do segmentacji budynków na lotniczych zdjęciach ukośnych. W tym przypadku okna i drzwi zostały potraktowane jako jedna klasa – „*opening area*". Oprócz tego uwzględnione były klasy takie jak ściana, dach oraz balkon. W innej pracy (Zhuo i in., 2019) autorzy zaś przedstawili metodę do semantycznej segmentacji budynków z obrazów UAV w oparciu o CNN. Istnieją również bardziej złożone metody analizy fasad budynków. Przykładowo w pracy (Mathias i in., 2016) zaproponowano metodę podzieloną na trzy etapy. W pierwszym kroku wykonywana jest inicjalna segmentacja semantyczna. W drugim wykorzystywane są detektory obiektów w celu poprawy początkowego etykietowania. Ostatni etap to uzupełnienie o dodatkową informację, która pozwala na elastyczne dostosowanie tej warstwy do różnych stylów elewacji. Dokładności na poziomie 85% uzyskano w pracy (Schmitz i Mayer, 2016), gdzie zaproponowano metodykę do semantycznej segmentacji fasad budynków bazując na sieci ConvNet. DeepFacade to kolejny przykład głębokiej konwolucyjnej sieci neuronowej (Liu i in., 2017; Liu i in., 2020) do segmentacji fasad budynków, w tym okien. Nowatorska koncepcja RFCNet oparta na głębokim uczeniu została zaproponowana przez Zhang i Aliaga (2022), którzy wykazali, że przewyższa ona inne modele takie, jak U-Net, Pix2Pix, EncNet, DeepLabv3+.

Jedną z ciekawych metod zaproponowano w artykule z 2023 roku (Mao i in., 2023), gdzie wykorzystano proces segmentacji budynków i naprawę szklanych fasad na fotogrametrycznych modelach 3D budynków poprzez wygładzanie siatki i mapowanie tekstur. W celu rozróżnienia szklanych fasad porównane było podejście segmentacji semantycznej ukośnych zdjęć lotniczych opartej na głębokim uczeniu oraz wyników segmentacji modeli budynków 3D. Analiza przeprowadzona wykazała, że dokładność wyniosła dla wskaźnika: *Precision* - 85% dla modelu U^2 -Net dla segmentacji zdjęć oraz 57% dla segmentacji modeli 3D metodą autorów oraz *Recall* - 90% dla modelu U^2 -Net dla segmentacji zdjęć oraz 50% dla segmentacji modeli 3D metodą autorów.

Oprócz szerokiego zastosowania DL do segmentacji obiektów takich jak okna na fasadach budynków, powstało również wiele metod wykorzystujących dane fotogrametryczne

do wykrywania obiektów małej architektury miejskiej takiej, jak latarnie, słupy energetyczne czy znaki drogowe.

Przykładowo, Mao i in., 2021 zaproponowali podejście do osadzania modeli znaków drogowych w oparciu o głębokie konwolucyjne sieci neuronowe, podejmując wyzwania związane ze słabą rekonstrukcją takich obiektów, osiągając wysoką dokładność (mAP równy 93.5%) wykrywania. W innym badaniu zaproponowano podejście do automatycznego wykrywania przydrożnych słupów energetycznych z obrazów Google Street View przy użyciu algorytmu RetinaNet (Zhang i in., 2018). Wykrywanie słupów energetycznych w mieście przy użyciu ortoobrazów z zastosowaniem sieci neuronowych takich jak Faster-RCNN, RetinaNet oraz Mask R-CNN zostało wykorzystane w pracach Gomes i in., 2020 oraz Chen i in., 2023.

Innym źródłem danych do wykrywania wysokich, smukłych obiektów a miastach mogą być chmury punktów. Istnieją aplikacje i rozwiązania, w których badacze wykorzystali modele do segmentacji chmur punktów w celu wykrycia latarni ulicznych (Shi i in., 2018, Huang i in., 2015; Aijazi i in., 2013).

Rozpoznawanie obiektów podobnych do słupów, takich jak latarnie uliczne, odbywa się także często za pomocą chmur punktów z mobilnego skanowania laserowego - MLS (Kang i in., 2018; Wang i in., 2021; Yan i in., 2016), ponieważ są one dobrze odwzorowane w takich danych. Innym dokładnym źródłem danych dla obiektów takich jak latarnie uliczne lub słupy wydają się być modele do wykrywania na obrazach UAV (Deokuliar, 2023; Chen i Miao, 2020). Przykładowym zbiorem danych jest TTPLA do wykrywania wież przesyłowych i linii energetycznych na zdjęciach z drona (Abdelfattah i in., 2020).

Chociaż wykrywanie małych obiektów jest nadal wyzwaniem (Kisantal i in., 2019), zapotrzebowanie na rozwiązania do rozpoznawania takich obiektów jest oczywiste, zwłaszcza jeśli chodzi o monitorowanie inteligentnych miast (Ahmed i in., 2022; Garg i in., 2021). Na podstawie powyższego przeglądu można wywnioskować, że ukośne zdjęcia lotnicze mogą dostarczyć wielu informacji geometrycznych, semantycznych, tekstur o wysokiej rozdzielczości i cech kontekstu przestrzennego, które są niezbędne do wykrywania obiektów takich jak okna czy latarnie uliczne w obszarach zurbanizowanych.

2.3. Ocena jakości i dostępności istniejących zbiorów danych uczących

Zbiory danych uczących są nieodzownym elementem rozwoju metod głębokiego uczenia, nie tylko w kontekście detekcji obiektów, ale też klasyfikacji czy segmentacji. Poprawnie utworzone zbiory danych odgrywają kluczową rolę w badaniach, gdyż służą jako podstawa do oceny wydajności i porównywania detektorów. Widoczna jest więc potrzeba pozyskiwania tego typu baz danych zbiorów uczących, a także publicznego udostępniania. Do tej pory powstało wiele zbiorów uczących, które można wykorzystywać do uczenia modeli sieci neuronowych. W tym miejscu należy wspomnieć o zbiorach uczących do detekcji na obrazach naturalnych, ponieważ stanowią największe bazy treningowe w widzeniu komputerowym. Pierwszym zbiorem danych wykorzystywanym przez badaczy był PASCAL Visual Object Classes (VOC). W późniejszym czasie opracowano zbiór danych ImageNet oraz MS COCO. Poniżej zestawiono podstawowe informacje o wymienionych zbiorach (Tabela 1).

Tabela 1 Podstawowe informacje o zbiorach danych do detekcji i segmentacji na obrazach naturalnych.

Zbiór danych	Liczba klas	Liczba zdjęć
PASCAL VOC (07++12)	20	22 591
MS COCO	80	330000
ImageNet	21 841	14197122

Jednakże odnosząc się do detekcji obiektów z wykorzystaniem danych fotogrametrycznych, w szczególności zdjęć lotniczych można powiedzieć, że wciąż brakuje zbioru danych przypominającego MS COCO czy ImageNet zarówno pod względem liczby obrazów, jak i liczby adnotacji. W ciągu ostatnich dziesięcioleci kilka różnych grup badawczych opublikowało zbiory danych w celu wykrywania obiektów na danych fotogrametrycznych i teledetekcyjnych. Ma to oczywiście ścisły związek z rozpowszechnieniem metod głębokiego uczenia w obserwacji Ziemi, ale także z zastosowaniami takimi jak autonomiczne samochody czy też w innych wysokopoziomowych zadaniach związanych z szeroko pojętym rozumieniem sceny.

Tabela 2 Zestawienie informacji o zbiorach uczących wykorzystywanych w dziedzinie obserwacji Ziemi do detekcji, segmentacji i klasyfikacji.

Nazwa zbioru	Liczba klas	Klasy	Dane	Liczba zdjęć	Liczba adnotacji	Rozdzielczość
TAS set	1	samochód	Google Earth	30	1319	brak inf.
VEDAI	9	pojazdy: łódź, samochód, samochód kempingowy, samolot,	zobrazowania satelitarne Utah AGRC	1210	3640	0.125 m

		pick-up, traktor, ciężarówka, van i inne				
UCAS-AOD	2	samochód, samolot	Google Earth	910	6029	0.3–2 m
DLR 3K Vehicle	1	samochód	zdjęcia lotnicze DLR 3K	20	14 235	0.13 m
AI-TOD v2	8	samolot, statek, samochód, zbiornik, most, basen, osoba, wiatrak	różne źródła zobrazowań satelitarnych, lotniczych, dronowych	28036	700621	0.3-30m
SODA-A	9	samolot, helikopter, mały samochód, duży samochód, statek, kontener, zbiornik, basen i wiatrak	Google Earth	2513	872069	brak inf.
FAIR1M	5 (37)	samolot, samochód, statek, boisko, droga	Gaofen, Google Earth	42796	1020579	0.3–0.8 m
RSOD	4	zbiornik, samolot, wiadukt, plac zabaw	Google Earth, Tianditu	976	6950	0.3–3 m
HRSC2016	1	statek	Google Earth	1070	2976	0.4–2 m
SZTAKI-INRIA dataset	1	budynek	QuickBird, Ikonos, Google Earth	9	665	0.5–1 m
DIOR	20	wiatrak, pojazd, stacja kolejowa, kort tenisowy, zbiornik, stadion, statek, port, boisko, pole golfowe, punkt poboru opłat na drodze ekspresowej, obszar obsługi drogi ekspresowej, tama, komin, most, wiadukt, boisko do koszykówki, boisko do baseballu, lotnisko, samolot	Google Earth	23463	192472	0.5-30 m
DOTA	18	samolot, statek, zbiornik, boisko baseballowe, kort tenisowy, basen, stadion, port, most, duży samochód, mały samochód, helikopter, rondo, boisko do piłki nożnej, boisko do koszykówki, suwnica bramowa, lotnisko i lądowisko dla helikopterów	Google Earth, Jilin-1, Gaofen-2, zdjęcia lotnicze	11268	1793658	0.1–4.5 m

xView	60	hangar dla samolotów, łódź, budynek, bus, samochód ciężarowy, samochód towarowy, betoniarka, suwnica, plac budowy, kontenerowiec, ciężarówka z dźwigiem, zniszczony budynek, wywrotka, koparka, pojazd techniczny, kuter rybacki, prom, spycharka, równiarka, ciężarówka, helikopter, samochód osobowy, lądowisko dla śmigłowców, namiot, lokomotywa, statek morski, żuraw samochodowy, łódź motorowa, tankowiec, pojazd osobowy, samolot pasażerski, płatowiec, obiekt, pick-up, słup, pojazd kolejowy, podnośnik, żaglówka, szopa, kontener wysyłkowy, wiele kontenerów, mały samolot, mały samochód, zbiornik, wózek widłowy, cysterna samochodowa, wieża, żuraw wieżowy, ciągnik, przyczepa, ciężarówka, wiele pojazdów, jacht, ciężarówka użytkowa, holownik, ciągnik, ciągnik z naczepą skrzyniową, ciągnik z naczepą płaską, ciągnik ze zbiornikiem cieczy	zobrazowania satelitarne WorldView-3	1413	1000000	0.3 m
NWPU VHR-10	10	samolot, statek, zbiornik, boisko do baseballu, kort tenisowy, boisko do koszykówki, stadion, port, most, pojazd	różne źródła zobrazowań satelitarnych, lotniczych	800	3651	0.3-2 m
AIR-SARShip-1.0	1	statek	Gaofen-3	31	3000	1 i 3 m
SIMD	15	samochód osobowy, samochód ciężarowy, samochód dostawczy, pojazd długi, autobus, samolot pasażerski, samolot śmigłowy, samolot szkoleniowy, samolot czarterowy, samolot myśliwski, inne, wózek widłowy, wózek spychający, helikopter, łódź	Google Earth	5000	45096	brak inf.
FGSD	43	różne rodzaje statków	Google Earth	2612	5634	0.12–1.93 m
COWC	1	samochód	różne źródła zobrazowań	53	32716	0.15 m
SSDD	1	statek	RadarSat-2, TerraSAR-X, and Sentinel-1	1160	2456	1-15 m
LEVIR	3	statek, samolot, zbiornik	Google Earth	22000	11000	0.2-1 m
HRRSD	13	samolot, boisko do baseballu, boisko do koszykówki, most, skrzyżowanie, boisko lekkoatletyczne, port, parking, statek, zbiornik, skrzyżowanie typu T, kort tenisowy, pojazd	Google Earth, Baidu Maps	21761	55740	0.15-1.2 m
MAR20	20	różne rodzaje samolotów	Google Earth	3824	22341	brak inf.

DOSR	20	różne rodzaje statków	Google Earth	1066	6172	0.5 - 2.5 m
ITCVD	1	samochód	zdjęcia lotnicze	228	23543	0.1 m
iSAID	15	samolot, statek, zbiornik, boisko do baseballu, kort tenisowy, boisko do koszykówki, stadion, port, most, duży pojazd, mały pojazd, helikopter, rondo, basen, boisko do piłki nożnej	różne źródła zobrazowań	2806	655451	brak inf.
RarePlanes	1	samolot	zobrazowania satelitarne WorldView3	50253	644258	0.3-1.5m
SpaceNet MVOI	1	budynek	zobrazowania satelitarne WorldView2	60000	126747	0.46-1.67 m
VisDrone	10	motor, bus, van, samochód, rower, osoba, pieszy, ciężarówka, wózek, trójkołowiec	UAV	10209	54200	brak inf.
CARPPK	1	samochód	UAV	1448	89777	brak inf.
Stanford Drone Dataset	6	piesi, deskorolkarze, rowerzyści, wózki, samochody i autobusy	UAV	929500	19000	brak inf.
UAVDT	1	saomchód	UAV	80000	841500	brak inf.
UAV20L, UAV123	3	człowiek, samochód, budynek	UAV	58700, 110000	brak inf.	brak inf.
SynDrone	28	droga, nawierzchnia, chodnik, pas drogowy, tory kolejowe, roślinność, teren, osoba, woda, samochód, ciężarówka, bus, motocykl, rower, pociąg, budynek, ściana, ogrodzenie, most, przeszkoda, słup, znak drogowy, światła uliczne, przeszkoda statyczna, przeszkoda poruszająca się, barierka	UAV	72000		brak inf.
MOHR	5	szkody powodziowe, budynek, samochód, ciężarówka, zapaść	UAV	10631	90014	brak inf.
AU-AIR	8	osoba, samochód, autobus, van, ciężarówka, rower, motocykl i przyczepa	UAV	32823	132034	brak inf.
UVSD	1	samochód	UAV	5874	98600	brak inf.

Powyższa Tabela 2 przedstawia przegląd wybranych zbiorów danych do wykrywania obiektów. Jak widać najbardziej popularne domeny dla opublikowanych zbiorów danych dotyczą wykrywania obiektów takich jak: samoloty, samochody, budynki, statki, pojazdy. Jednocześnie należy podkreślić, że kluczowym aspektem jest to, że większość zbiorów cechuje się niską różnorodnością kategorii obiektów, co stanowi ograniczenie do wykorzystywania takiego zbioru do niektórych zastosowań. Przykładem takiego problematycznego zagadnienia może być wykrywanie obiektów architektury miejskiej, które są głównym przedmiotem badań w pracy. Powyższy przegląd pokazuje zestawienie wielu zbiorów uczących dla danych UAV,

zdjęć lotniczych czy scen satelitarnych. Kompleksowy przegląd ponad 400 publicznie udostępnionych zbiorów danych w różnego typu aplikacjach dotyczących obserwacji Ziemi można znaleźć w artykule przeglądowym opublikowanym przez badaczy w 2022 roku (Xiong i in., 2022). Szczegółowe informacje o 91 zbiorach danych do detekcji obiektów zostały zestawione w tabeli z podziałem na 18 różnych obszarów badawczych. W artykule (Ding i in., 2021), w którym przedstawiono zestaw danych do wykrywania obiektów na obrazach lotniczych (DOTA), który zawiera 1 793 658 instancji obiektów z 18 kategorii zebranych w oparciu o 11 268 zdjęć lotniczych, również porównano inne istniejące zbiory danych do wykrywania obiektów na zdjęciach lotniczych. DOTA aktualnie jest największym zbiorem danych do wykrywania obiektów w badaniu Ziemi. Jednak należy zaznaczyć, że w tym zbiorze wśród 18 klas również nie uwzględnione zostały obiekty małej architektury.

Powyższy przegląd uwydatnia problem z dostępnością zbiorów uczących do detekcji obiektów, w szczególności na lotniczych zobrazowaniach ukośnych. Powyższe zestawienie potwierdza również, że otwarte źródłowo zbiory dotyczą w większości typowych obiektów lub scenerii, głównie za pomocą rzutów z góry zamiast ujęć ukośnych. W dalszej części pracy porównane zostaną różne źródła zobrazowań uwzględniające potencjał wykorzystania do detekcji obiektów małej architektury miejskiej.

2.3.1. Źródła obrazów do detekcji obiektów miejskich

Analizując przestrzeń miejską, w szczególności gdy mowa jest o obszarach szybko rozwijających i zmieniających się, widoczne jest zapotrzebowanie na dokładne dane pozyskiwane na tyle często, aby aktualizacja geoprzestrzennych baz danych była wydajna i wykonywana regularnie. W ramach prac badawczych zdecydowano skupić się na zadaniu detekcji, a dokładniej na wykrywaniu obiektów małej architektury i automatycznym wyodrębnianiu informacji o nich ze zdjęć.

Obiektami małej architektury miejskiej z definicji[3] są drobne elementy architektoniczne, którymi w mieście wypełniona jest przestrzeń. Zalicza się do tej kategorii m.in. latarnie, kapliczki, altany, ogrodzenia, huśtawki, place zabaw, kosze na śmieci, uliczne ławki, wiaty przystankowe, stojaki rowerowe, słupy ogłoszeniowe. Widać więc, że są to elementy krajobrazu, które pełnią w przestrzeni miasta niezbędną rolę. Podążając więc w kierunku

[3] https://promarlighting.pl/co-jest-mala-architektura/

inteligentnych miast konieczne jest zasilanie baz danych w miastach informacjami o takich obiektach np. w celu budowania szczegółowych modeli 3D miast czy wspomagania monitorowania i zarządzania przestrzenią miejską. Jedną z baz danych w Polsce, w której gromadzi się informacje dotyczące obiektów małej architektury jest baza danych obiektów topograficznych BDOT500.

Jednakże jak zostało wspomniane ekstrakcja małych obiektów miejskich stanowi wyzwanie, a główny problem jest związany z niewielkim rozmiarem w stosunku do rozmiaru całego obrazu (Yang i in., 2019). Dlatego w przypadku takich obiektów należy zastanowić się, który typ zobrazowań fotogrametrycznych (sceny satelitarne, zobrazowania lotnicze, zdjęcia UAV) będzie najlepszym źródłem danych. Dobór technologii do konkretnego typu zastosowania powinien uwzględniać zarówno warunki dokładnościowe, jak i czynniki ekonomiczne.

Wracając do zestawienia (Tabela 2) wykonanego na potrzeby pracy, znalazły się również w nim zbiory satelitarne. W przypadku tego typu danych naturalnym ograniczeniem jest rozdzielczość przestrzenna, gdy mowa o detekcji takich obiektów jak znaki drogowe, studzienki czy słupy reklamowe.

Aktualnie zobrazowania ukośne (zarówno lotnicze, jak i z drona) są jednym z najważniejszych źródeł danych do zastosowań miejskich. Zawierają informacje o górnej części i bokach obiektów. Porównując zdjęcia ukośne z obrazami naziemnymi i pionowymi, można zauważyć, że w zobrazowaniach ukośnych występują większe zniekształcenia perspektywy. Jednocześnie potencjał, jaki niesie ze sobą fotogrametria ukośna jest dość znaczący. Obiekty na zdjęciach ukośnych są fotografowane z różnych perspektyw w różnej skali, poza tym są odwzorowane kilkukrotnie na zdjęciach wykonanych z różnych kierunków w odstępach czasowych. Biorąc pod uwagę ukośne zdjęcia należy zastanowić się nad wadami i zaletami zdjęć UAV i zdjęć lotniczych.

Najczęściej wykorzystywanymi zbiorami danych do modelowania 3D w skali miasta i detekcji obiektów zainteresowania są zdjęcia ukośne UAV (Yang i in., 2021). Wydaje się więc, że zdjęcia ukośne z drona byłyby dobrym rozwiązaniem do detekcji obiektów małej architektury miejskiej. Oczywiście dane UAV wykazują duży potencjał w analizie takich scen. W ostatnich latach zostały poczynione badania pokazujące zastosowanie danych UAV do zadań monitorowania ruchu na drodze (Huang i in., 2021), analizy roślinności miejskiej (Lee i in., 2021), wyznaczania znaków drogowych (Guan i in., 2022), monitorowania środowiska,

zarządzania i tworzenia planów zagospodarowania przestrzennego (Muhmad Kamarulzaman i in., 2023).

Przykładów zastosowań danych z dronów można wymieniać bardzo dużo. Co więcej, prawdziwe wydaje się stwierdzenie, że obrazowanie z dronów zrewolucjonizowało gromadzenie danych, wprowadzając technologię do różnych dziedzin. Dodatkowo, zobrazowania UAV oferują możliwość pozyskiwania danych w czasie zbliżonym do rzeczywistego, co jest przydatne przy inspekcji szkód czy monitorowania smogu w ośrodkach miejskich.

Chociaż zastosowanie danych UAV zapewnia odpowiednią rozdzielczość dla zadań takich, jak detekcja obiektów małej architektury, to mają one pewne ograniczenia. Zebranie dokładnych danych UAV wymaga często długiego czasu lotu, aby pokryć duży obszar miasta.

Jeszcze jakiś czas temu można było powiedzieć, że ze względu na większą wysokość, na której latają samoloty (względem dronów), rozdzielczość i szczegółowość zdjęć lotniczych będzie miała trudności z dopasowaniem się do wymagań potrzebnych do precyzyjnych zastosowań. Jednakże najnowsze systemy takie jak Leica CityMapper-2 umożliwiają tworzenie szczegółowej warstwy informacyjnej 3D. Więcej informacji o sensorze zostało opisane w rozdziale 4.2. W tym miejscu należy jednak nadmienić, że jest to tak zwany „*state-of-the-art*" jeśli chodzi o rozwiązania lotnicze na rynku. Sensor ten został specjalnie zaprojektowany do mapowania obszarów miejskich z wysoką wydajnością. Dzięki zastosowanej technologii (obrazowanie rejestrujące w kierunku pionowym i w czterech kierunkach ukośnych wraz z synchronicznie pozyskanymi danymi LiDAR) zapewniona jest najwyższa rozdzielczość do wizualizacji każdej części miasta, przez co wykrywanie takich obiektów jak latarnie miejskie czy słupy reklamowe może być dużo łatwiejsze i wydajne.

Biorąc pod uwagę wymienione aspekty zdecydowano, że do części eksperymentalnej zostaną wykorzystane lotnicze zdjęcia ukośne i będą głównym przedmiotem badań.

2.3.2. Detekcja obiektów małej architektury miejskiej – dostępność zbiorów uczących

Pomimo coraz większego zainteresowania detekcją obiektów miejskich na zdjęciach, a co za tym idzie powstawaniem nowych zestawów treningowych dostępność zbiorów uczących dla obiektów małej architektury miejskiej wciąż jest bardzo niska. Potwierdzeniem tego jest wynik przeprowadzonej analizy literatury pod względem dostępności zbiorów

uczących dla tych obiektów. W Tabela 3 zestawiono otwarte źródłowo przykładowe zbiory danych dla wybranych obiektów małej architektury miejskiej. Podzielono zbiory uczące na dwie grupy: zdjęcia naziemne (naturalne sceny) oraz zobrazowania fotogrametryczne (w tym zdjęcia lotnicze, satelitarne i UAV).

Tabela 3 Zestawienie zbiorów uczących dla wyodrębnionych obiektów małej architektury miejskiej wykonane na podstawie przeglądu literatury.

Obiekt	Zbiór uczący – zdjęcia naziemne	Zbiór uczący – zdjęcia lotnicze, satelitarne, UAV
Ławki	Mapillary Vistas Dataset, MS COCO	-
Reklamy	Mapillary Vistas Dataset, MS COCO	-
Latarnie	Mapillary Vistas Dataset, MS COCO	Urban Road Facilities Dataset
Znaki drogowe	Mapillary Vistas Dataset, Cityscapes Dataset, GTSRB, CURE-TSR, CURE-TSD, Tsinghua-Tencent 100K	Zbiór uczący z pracy (Mao i in., 2021)
Słupy (uliczne, energetyczne)	MS COCO, Mapillary Vistas Dataset	Urban Road Facilities Dataset
Parkometry	MS COCO	-
Ogrodzenia	Mapillary Vistas Dataset, MS COCO, Cityscapes Dataset	Semantic Drone Dataset
Studzienki	Mapillary Vistas Dataset, Storm-drain and manhole detection,	Prades-le-Lez database, Gigean dataset
Kapliczki, altany	-	-

Jak widać w powyższej tabeli dla niektórych obiektów nie istnieją żadne otwarte źródłowo zbiory danych do detekcji na danych fotogrametrycznych. Spośród analizowanych obiektów małej architektury miejskiej większym zainteresowaniem cieszą się obiekty infrastruktury drogowej, dla których został na przykład utworzony zbiór danych Urban Road Facilities Dataset (Mao i in., 2023). Obejmuje on 1075 obrazów z adnotacjami dla różnych miejskich obiektów drogowych, podzielonych na osiem kategorii, takich jak różne typy latarni ulicznych, urządzeń monitorujących i sygnalizacji świetlnej. Zbiór danych został utworzony przy użyciu obrazów z drona DJI M600 (GSD wynosi 2-6 cm).

W zestawieniu nie zostały uwzględnione chmury punktów, które też często są wykorzystywane do wykrywania tego typu obiektów. Zarówno metody wykorzystujące dane z mobilnego skanowania laserowego (MLS), jak i chmury TLS (naziemny skaning laserowy), ULS (chmury

punktów z drona) czy ALS (lotniczy skaning laserowy). Przedmiotem badań w pracy są głównie lotnicze zdjęcia ukośne, dlatego ograniczono się do przeglądu zbiorów uczących dla tego typu danych.

Podsumowując część dotyczącą oceny jakości i dostępności istniejących zbiorów danych uczących, podkreślono niedobór zestawów uczących dla małych obiektów architektury miejskiej i szersze zapotrzebowanie na bardziej kompleksowe zestawy treningowe w tym obszarze. Zaakcentowana luka badawcza w wykorzystaniu danych fotogrametrycznych do wykrywania obiektów miejskich jest obszarem zainteresowania niniejszej pracy doktorskiej.

2.4. Żonglowanie reprezentacjami - transfer informacji między danymi fotogrametrycznymi

Niezależnie od typu danych, jeśli mają nadaną georeferencję, możliwy jest transfer cech i atrybutów pomiędzy różnymi reprezentacjami obiektów. Wykorzystując zależności matematyczne można w łatwy sposób przenosić np. informacje semantyczne bądź etykiety między chmurami punktów, zdjęciami czy modelem mesh 3D. Zatem w odniesieniu do metod DL możliwa jest redukcja wysiłku związanego z ręcznym tworzeniem zbioru uczącego. Zadanie sprowadza się wtedy do etykietowania jednej reprezentacji danych (np. tylko na zdjęciach), a potem przenosi się na inny produkt (np. na model mesh 3D dodając informację semantyczną). Taki proces transferu informacji w publikacji Laupheimer i Haalaa (2022) został określony jako żonglowanie reprezentacjami (*ang. juggling with representations*).

2.4.1. Automatyzacja tworzenia zbiorów uczących

Rozwój metod DL wymagających dużej ilości danych treningowych wykazuje jednocześnie potrzebę wydajnych strategii generowania takich baz. Chociaż w przypadku niektórych zadań duże zbiory danych z adnotacjami są dostępne publicznie, to w niektórych obszarach brak zbiorów treningowych i testowych stanowi wyzwanie. Oprócz wspomnianych metod, które wspierają problem niedostępności zbiorów treningowych tj. *data augmentation* oraz *transfer learning*, stosuje się również inne metody automatyzacji procesu generowania adnotacji, zarówno dla danych obrazowych, jak i 3D. Tego typu rozwiązania są stosowane zarówno w segmentacji obrazów, jak i wykrywaniu obiektów.

Zhuo i in., (2018) zaproponowali automatyczne generowanie adnotacji na obrazach. Metoda składa się z trzech kroków: ręcznego etykietowanie jednego lub dwóch obrazów lotniczych;

przeniesienia etykiet pikseli na wiele obrazów UAV za pośrednictwem chmury punktów UAV; oraz udoskonalenia wygenerowanych adnotacji przy użyciu gęsto sprzężonego modelu CRF (*ang. Conditional Random Fields)* i naiwnego klasyfikatora Bayesa, który przybliża prawdopodobieństwo przynależności obserwacji do poszczególnych klas. Kolejną propozycją (Xie i in., 2016) transferu informacji między danymi jest wykorzystanie poetykietowanych danych LiDAR lub modeli 3D w celu przeniesienia adnotacji, w tym przypadku instancji semantycznych z 2D do 3D. W innej pracy (Braun i Borrmann, 2019) przedstawiono nowatorską metodę automatycznego etykietowania zdjęć opartą na połączeniu modelu BIM (*ang. Building Information Model*) i chmur punktów, rzutując elementy obiektu budowlanego na obrazy w celu uzyskania informacji semantycznej.

Automatyzacja etykietowania danych z wykorzystaniem modelu SAM została wykorzystana w pracy Gallagher i in., 2024, gdzie dane do wykrywania obiektów na obrazach wielospektralnych zostały utworzone na podstawie adnotacji na obrazach RGB. Podobna metoda automatycznego generowania semantycznych adnotacji segmentacyjnych dla obrazów termalnych UAV, bazując na danych satelitarnych została zaproponowana w pracy Lee i in., 2024.

2.4.2. Dane syntetyczne

Innym rozwiązaniem, które stosuje się w literaturze aby sprostać problemom braku baz treningowych jest wykorzystywanie danych syntetycznych do wstępnego trenowania sieci neuronowych. W swojej pracy Ros i in. (2016) zaproponowali generowanie syntetycznych obrazów z adnotacjami na poziomie pikseli, tworząc syntetyczny zbiór danych scen miejskich o nazwie SYNTHIA do segmentacji 11 klas. Syntetycznie wygenerowane dane referencyjne zostały przez Griffiths i Boehm (2019). SynthCity to otwarty zbiór danych składający się z syntetycznej chmury punktów mobilnego skanowania laserowego przeznaczony do klasyfikacji/segmentacji chmur w dziewięciu kategoriach typowych dla środowiska miejskiego. Eksperymenty przeprowadzone na syntetycznych danych STPLS3D również potwierdzają skuteczność zastosowania wygenerowanych w ten sposób chmur punktów 3D (Chen i in., 2022).

Chociaż dane syntetyczne wydają się być ciekawym rozwiązaniem podczas, gdy niemożliwe jest pozyskanie rzeczywistych danych, to mają swoje ograniczenia. W przypadku tworzenia danych syntetycznych można oczywiście zadbać o zrównoważony rozkład klas, wykluczyć

niską jakość próbek danych, jednak konieczne jest zapewnienie danych podobnej jakości jak dane realne. Gdy model DL wytrenowany zostanie na sztucznie wygenerowanych chmurach punktów, które są przykładowo pozbawione szumów, a następnie predykcja zostanie przeprowadzona na surowych danych, to wynik może okazać się dużo słabszy. Zapewne w niektórych aplikacjach rozwiązanie z tworzeniem sztucznych danych może być wystarczające, jednak należy rozważyć czy automatyzacja manualnej pracy związanej z etykietowaniem nie przyniosłaby lepszych efektów w konkretnym zadaniu.

Część eksperymentalna

3. Plan eksperymentów

Pogłębiona analiza literatury i wyniki przeprowadzonych badań wskazały na problemy badawcze, które zostały przybliżone na samym początku pracy. Ponadto jak zostało wspomniane przyjęta metodyka zakłada wykonanie eksperymentów celem odpowiedzi na kolejne pytania badawcze. Poniżej opisany został plan badań, który został zaprojektowany zgodnie z założoną metodyką dla osiągnięcia postawionych celów i udowodnienia postawionej na wstępie naukowej tezy.

Pierwsza część dotyczyła wyboru obiektów do zadania detekcji. Zdecydowano, że będą nimi latarnie uliczne oraz okna jako dobre przykłady reprezentatywne w kontekście przestrzeni miejskiej. Następnie dokonano doboru danych do testów. Pożądanym źródłem były lotnicze zdjęcia ukośne, dlatego też zdecydowano, że do eksperymentów wykorzystane zostaną dane hybrydowe pozyskane sensorem Leica CityMapper2. Przedstawione zostały możliwości wykorzystania w odniesieniu do problemów i celów badawczych postawionych w pracy.

W pierwszym zaprojektowanym eksperymencie przeanalizowane zostały chmury punktów z gęstego dopasowania zdjęć lotniczych w celu zbadania kompletności chmur punktów względem referencyjnych danych z lotniczego skanowania laserowego w odniesieniu do reprezentacji latarni ulicznych i przydatności do tworzenia zbiorów uczących w sposób zautomatyzowany.

W tej części wykorzystano opracowaną wcześniej metodykę (Zachar i in., 2022), w której wykorzystuje się fragmenty wyciętych chmur punktów dla obiektów zainteresowania jako dane wejściowe do tworzenia zbioru treningowego do detekcji na zdjęciach ukośnych. Współrzędne chmury punktów z układu terenowego przeliczone zostały do układu pikselowego zdjęcia i w ten sposób powstał zbiór uczący składający się ze zdjęć i informacji o ramkach ograniczających.

Powstały zbiór uczący do detekcji latarni na zdjęciach ukośnych, który był wynikiem pierwszego etapu badań posłużył do treningu sieci neuronowej YOLO. Celem tego eksperymentu było zweryfikowanie potencjału wykorzystania metod DL do wykrywania latarni na lotniczych zdjęciach ukośnych. Mając na uwadze fakt, że innym podejściem do wykrywania obiektów na zdjęciach jest segmentacja semantyczna, utworzony zestaw danych został wykorzystany w kolejnym eksperymencie jako dane referencyjne do porównania

wyników wykrywania przy użyciu modelu Segment Anything Model (SAM). Przeprowadzone na tym etapie prace umożliwiły sprawdzenie, czy taki gotowy, wstępnie wytrenowany model może obsługiwać wykrywanie latarni bez potrzeby trenowania sieci od podstaw. Ponadto jest to odpowiedni przykład tego, jak można wykorzystywać różne podejścia do automatyzacji tworzenia zbioru uczącego w podejściu nienadzorowanym.

Na podstawie wyników uzyskanych dla latarni ulicznych, które potwierdziły potencjał stosowania lotniczych zdjęć ukośnych w uczeniu głębokim zaprojektowano kolejny eksperyment, który dotyczył drugiego analizowanego obiektu. Podjęta została próba wykrywania okien na innych obszarach badawczych z wykorzystaniem danych o tej samej charakterystyce realizując semantyczną metodę segmentacji. Ponownie wykorzystano model SAM, jednak do obiektu o innych cechach.

Ostatnią częścią przeprowadzonych eksperymentów było wykorzystanie wyników detekcji na zdjęciach w celu wyznaczenia położenia obiektu w terenowym układzie współrzędnych. Informacja o lokalizacji jest ważna w aspekcie tworzenia oraz aktualizacji baz danych o obiektach miejskich. Oczywiście istnieje wiele możliwości w jaki sposób można podejść do tematu, jednak kluczowe jest wykorzystanie informacji o elementach orientacji wewnętrznej i zewnętrznej zdjęć oraz tego, że obiekt został wykryty na wielu zdjęciach z kilku kierunków. W tej części badań skupiono się zatem na opracowaniu metodyki łączenia wyników detekcji z wielu kierunków, tak by wyznaczone współrzędne w układzie terenowym były wystarczająco dokładne dla potrzeb baz danych o obiektach miejskich. Ponadto mając te same dane źródłowe do detekcji (w tym przypadku lotnicze zdjęcia ukośne), ale innego typu obiekty należało opracować dwa różne podejścia, oddzielnie dla okien i latarni. Z jednej strony metodyka odnosiła się do łączenia wyników detekcji dla latarni, gdzie wynikiem z modelu były ramki ograniczające. Natomiast drugie podejście dotyczyło łączenia wyników z wielu kierunków dla okien, dla których wynikiem modelu były segmenty, a dokładniej poligony dokładnie otaczające obiekt.

4. Charakterystyka wykorzystanych danych

4.1. Dane hybrydowe

Analizując dostępność i jakość danych geoprzestrzennych dla miast oraz kierunki rozwoju, w którym idą inteligentne miasta, zdecydowano, że na potrzeby prac badawczych wykorzystane zostaną dane hybrydowe. Hybrydowe systemy mapujące, które zbierają synchronicznie dane LiDAR oraz zdjęcia pionowe i ukośne staną się prawdopodobnie standardem, jeśli chodzi o pozyskiwanie danych w miastach (Toschi i in., 2018; Bacher, 2021). Ponadto mogą zapewnić wysoką kompletność produktów, o jednoczesnej wysokiej jakości geometrycznej, co jest pożądane nie tylko jeśli chodzi o przetwarzanie danych, ale również wykrywanie obiektów (Toschi i in., 2019). Koncepcja zaproponowana przez Bachera (2021) w odniesieniu do obszarów zurbanizowanych mówi o tym, że w przypadku zdjęć (w domyśle pionowych i ukośnych) GSD powinno wynosić 3-10 cm, a chmury punktów LiDAR powinny charakteryzować się gęstością powyżej 15 punktów/m2.

4.2. Znaczenie lotniczych zdjęć ukośnych

- *Charakterystyka lotniczych zdjęć ukośnych*

Lotnicze zdjęcia ukośne są definiowane jako zdjęcia wykonywane przy wychyleniu osi optycznej kamery od pionu pod kątem 30-45 stopni. Zalety fotogrametrii ukośnej obejmują możliwość pokazania fasad budynków i obiektów w widoku perspektywistycznym, których nie widać na zdjęciach pionowych w całości. W rezultacie zdjęcia ukośne można powiedzieć, że są łatwiejsze do zrozumienia, gdyż przypominają obraz rzeczywisty.

W przeciwieństwie do lotniczych zdjęć pionowych, ukośne zdjęcia mają pewne właściwości, które należy wziąć pod uwagę przy rozważaniu geometrii mapowania. Chodzi dokładniej o skalę obrazowania, która w obrębie zdjęcia ukośnego będzie zmienna. Sprowadza się to do tego, że rozmiar piksela (*ang. Ground Sampling Distance - GSD*) zmienia się znacząco wzdłuż obrazu. Dokładniej na pierwszym planie zdjęcia skala jest większa, a GSD mniejsze, gdy równocześnie mniejsza skala i większe GSD jest w tle obrazu. Należy pamiętać o tych czynnikach podczas planowania bloku zdjęć ukośnych, jak i podczas opracowywania takiego bloku.

W przeszłości zobrazowania ukośne były wykorzystywane raczej do wizualizacji i interpretacji niż do zastosowań metrycznych. Obecnie natomiast w fotogrametrii łączy się tradycyjne zdjęcia pionowe z ukośnymi na przykład w celu tworzenia trójwymiarowych modeli miast.

Fotogrametria ukośna zapewnia znacznie bardziej szczegółowe i wiarygodne pomiary, zwłaszcza w pracy z modelami 3D obszarów miejskich (Gajski i Dzięgielewska-Gajski, 2023).

- *Charakterystyka sensora Leica CityMapper-2*

Ze względu na to, że w pracy oba zestawy zdjęć ukośnych pozyskane zostały sensorem Leica CityMapper-2 zdecydowano się na przybliżenie parametrów i podkreślenie zalet płynących z wykorzystania w celach analizy miast.

Leica CityMapper-2 to jedno z najlepszych rozwiązań lotniczych na rynku przeznaczony do tworzenia szczegółowej warstwy informacyjnej, w tym ortofotomapy, chmur punktów, modeli budynków 3D. Głównym założeniem systemu było zwiększenie wydajności zbierania danych 3D w obszarach miejskich. Dzięki czemu może być wykorzystywany w efektywny sposób do tworzenia cyfrowych bliniaków 3D miast. Sensor jest hybrydowym rozwiązaniem, który składa się z systemu optycznego oraz sensora LiDAR. Rejestrowane są podczas nalotu dwa obrazy nadirowe (RGB/NIR) oraz cztery obrazy ukośne. Kąt wychylenia osi optycznej dla kamer ukośnych wynosi 45°. System optyczny stanowią kamery wyposażone w matrycę CMOS o rozdzielczości 150 Mpx i unikalną mechaniczną kompensację ruchu. W przeciwieństwie do tradycyjnych systemów kamer, takie rozwiązanie zapewnia możliwość wykonywania zdjęć w trudnych warunkach oświetleniowych bez zmniejszania prędkości lotu. Częstotliwość skanera LiDAR - Leica Hyperion2+, który składa się na system wynosi 2 MHz. Można powiedzieć więc, że dane pozyskane tego typu sensorem są najbardziej dokładnymi danymi, jakie współcześnie można pozyskać z poziomu lotniczego.

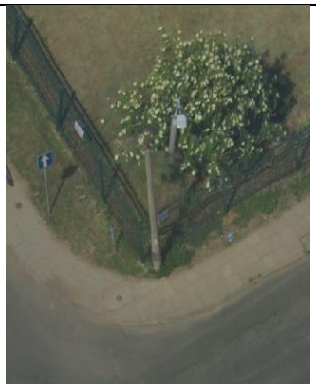
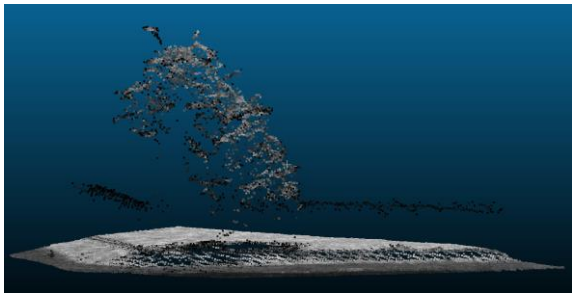
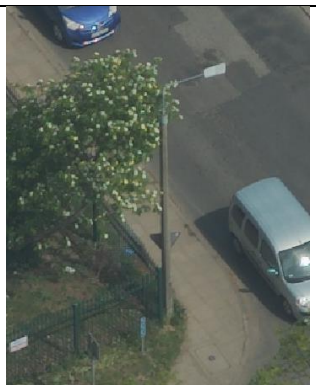
5. Obiekty badawcze

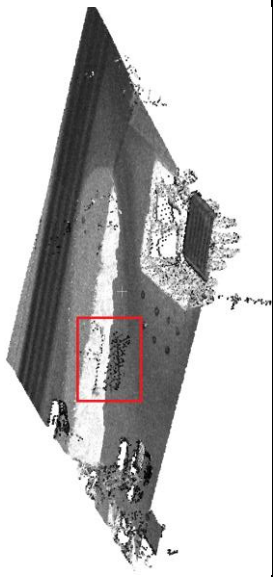
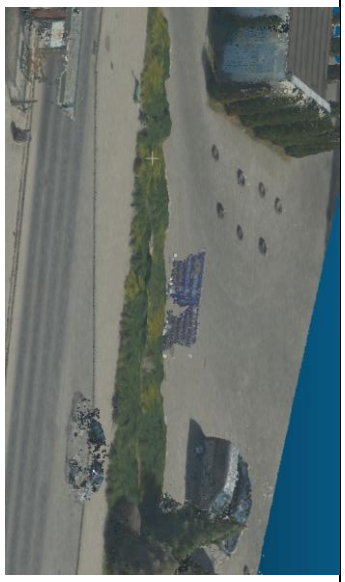
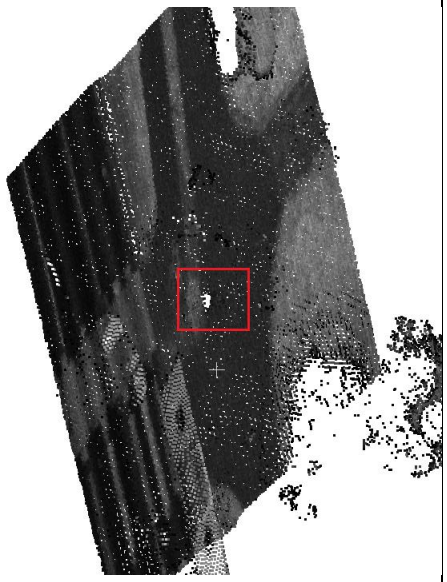
Jak już nakreślono, analiza przestrzeni miejskiej cieszy się coraz większym zainteresowaniem wielu środowisk badawczych. Według Organizacji Narodów Zjednoczonych, do 2050 r. ok. 68% populacji ma zamieszkiwać miasta. W związku z tym pojawiają się nowe wyzwania w zarządzaniu przestrzenią miejską. Widoczna jest potrzeba pozyskiwania nowych danych przestrzennych i ich dalszej interpretacji w celu wydobycia informacji i wiedzy. Nieodłącznym elementem krajobrazu miejskiego stanowią budynki i ich elementy takie jak fasady, okna, balkony, drzwi. Rozmieszczenie i charakterystyki wymienionych elementów mogą dostarczyć informacji na temat struktury budynków oraz ogólnego układu urbanistycznego. Wraz z rozwojem sensorów do zbierania danych 3D i technologii AI możliwa jest skuteczna detekcja na przykład na lotniczych zdjęciach ukośnych. Wykrywanie i rozróżnianie elementów budynków na tego typu danych niesie za sobą wiele korzyści. Ukośne zdjęcia lotnicze, jak już wspomniano są szeroko stosowanymi do modelowania 3D scen miejskich na dużą skalę, zapewniając jednocześnie informacje geometryczne. Ponadto charakterystyka zdjęć ukośnych pozwala na to, że elementy budynków, szczególnie w kontekście ścian są dobrze odwzorowane na tego typu danych.

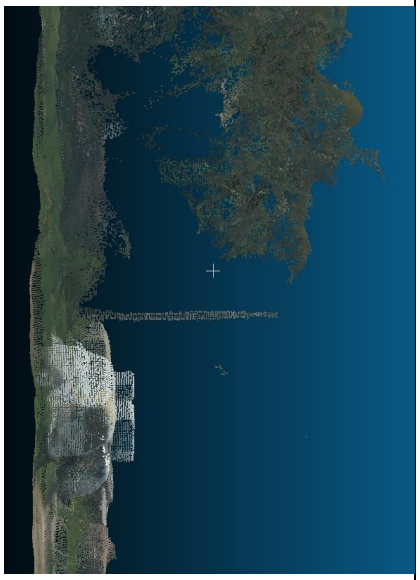
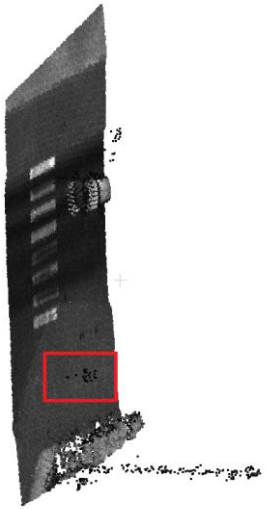
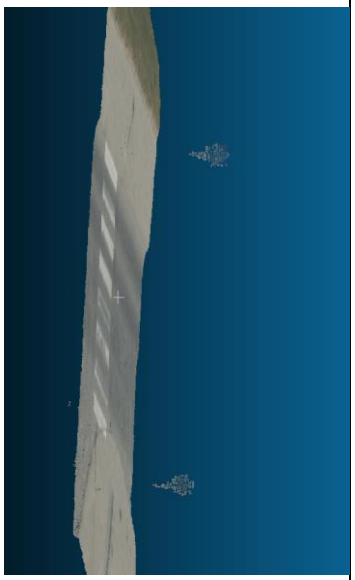
Obiektem, który zdecydowano się wykrywać w ramach badań są okna. Pomimo, że okna wydają się być bardzo łatwym obiektem, jeśli chodzi o ich interpretację, to istnieje parę wyzwań związanych z definicją okna. Większość okien ma symetryczny i prostokątny kształt, jednak niekoniecznie ich detekcja jest trywialnym zadaniem. Przykładowo w przypadku gdy na danym obszarze występują okna z okiennicami należy zastanowić się czy powinny być wykrywane razem, czy należy rozróżnić takie obiekty na poszczególne typy. Ponadto różnorodność tych obiektów jest na tyle duża, że detektor wytrenowany na danych z jednego obszaru może sobie nie poradzić z detekcją okien dla miasta o innym charakterze.

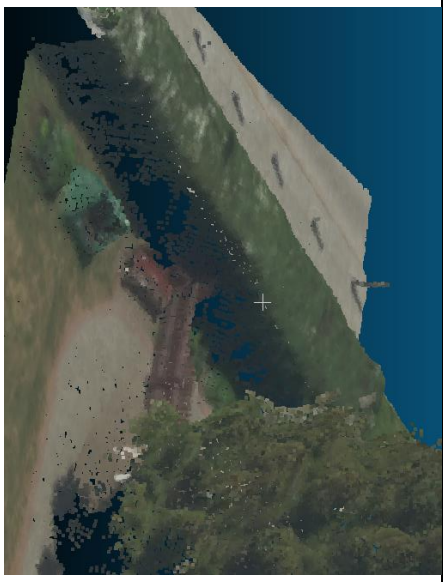
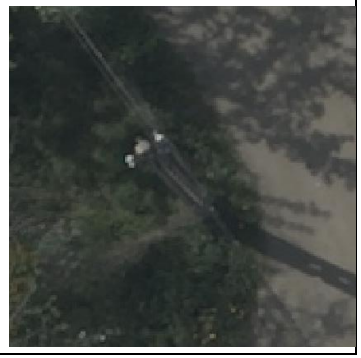
Oprócz okien zdecydowano również na badania z wykorzystaniem innego obiektu badawczego spośród małej architektury miejskiej. Wybór obiektów do części eksperymentalnej był związany z odwzorowaniem na różnego typu danych, gdyż proponowana metodyka zakładała automatyzację procesu tworzenia zbiorów uczących z wykorzystaniem chmur punktów. Brane były pod uwagę chmury punktów DIM (*ang. dense image matching*) oraz ALS (*ang. airborne laser scanning*). Przeanalizowano różne obiekty i ich widoczność na danych fotogrametrycznych. Poniższa Tabela 4 przedstawia wynik odwzorowania różnych obiektów miejskich na danych DIM, ALS oraz ortofotomapie.

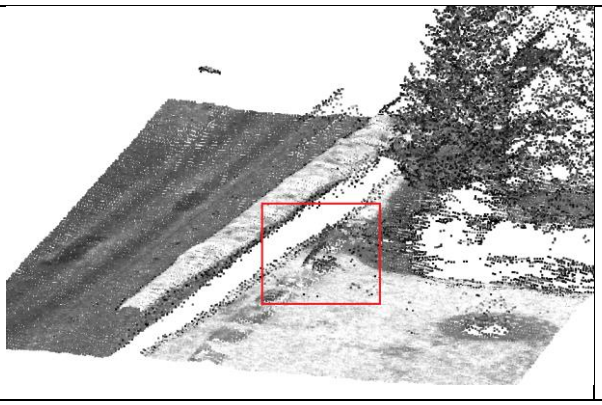
Tabela 4 Porównanie odwzorowania wybranych obiektów architektury miejskiej na różnych typach danych (chmurach DIM, ALS, zdjęciach pionowych i ukośnych oraz ortofotomapie).

LATARNIE ULICZNE	CHMURA PUNKTÓW DIM		ORTOFOTOMAPA	
	CHMURA PUNKTÓW ALS		ZDJĘCIE PIONOWE	
			ZDJĘCIE UKOŚNE	
ŚMIETNIK	CHMURA PUNKTÓW DIM		ORTOFOTOMAPA	

<table>
<tr><td colspan="3">BANER REKLAMOWY</td><td colspan="2"></td></tr>
<tr><td colspan="2">CHMURA PUNKTÓW ALS</td><td>CHMURA PUNKTÓW DIM</td><td colspan="2">CHMURA PUNKTÓW ALS</td></tr>
<tr><td colspan="2">  </td><td>  </td><td colspan="2">  </td></tr>
<tr><td>ZDJĘCIE UKOŚNE</td><td>ZDJĘCIE PIONOWE</td><td>ORTOFOTOMAPA</td><td>ZDJĘCIE UKOŚNE</td><td>ZDJĘCIE PIONOWE</td></tr>
<tr><td>  </td><td>  </td><td>  </td><td>  </td><td>  </td></tr>
</table>

SŁUP ENERGETYCZNY	ZNAK DROGOWY		
CHMURA PUNKTÓW DIM	CHMURA PUNKTÓW ALS		CHMURA PUNKTÓW DIM
			
ORTOFOTOMAPA	ZDJĘCIE UKOŚNE	ZDJĘCIE PIONOWE	ORTOFOTOMAPA
			

<table>
<tr><td colspan="2">ŁAWKA</td><td colspan="2"></td></tr>
<tr><td>CHMURA PUNKTÓW ALS</td><td>CHMURA PUNKTÓW DIM</td><td colspan="2">CHMURA PUNKTÓW ALS</td></tr>
<tr><td></td><td>  </td><td colspan="2">  </td></tr>
<tr><td>ZDJĘCIE PIONOWE</td><td>ORTOFOTOMAPA</td><td>ZDJĘCIE UKOŚNE</td><td>ZDJĘCIE PIONOWE</td></tr>
<tr><td>  </td><td>  </td><td>  </td><td>  </td></tr>
</table>

			ZDJĘCIE UKOŚNE	

Analizując powyższy przegląd, widać, że problemy występują w przypadku chmur punktów z gęstego dopasowania obrazów. Zdarza się, że obiekty nie są w ogóle widoczne w chmurze punktów, bądź występują pewne braki i niektóre elementy obiektu nie zostały w pełni odwzorowane. W odniesieniu do chmur punktów z dopasowania, jeśli obiekt znajduje się wśród drzew, również mogą wystąpić problemy z odwzorowaniem. Penetracja roślinności jest cechą skanowania laserowego i takie rozwiązanie częściowo może rozwiązać problemy DIM. Natomiast stosując tego typu dane do wykrywania obiektów czy też podczas wsparcia w tworzeniu zestawu treningowego, ważne jest, aby obiekty te były w pełni odwzorowane w danych.

W odniesieniu do obiektów architektury miejskiej ciekawą grupą, znacząco inną niż okna są wysokie, smukłe elementy wystające ponad powierzchnię ziemi takie, jak latarnie czy słupy energetyczne. Chociaż jak widać w tabeli latarnie uliczne są odwzorowane na ortofotomapach, to ich wykrywanie na takim produkcie może być trudne. Wynika to z niedoskonałości oraz deformacji i zniekształceń występujących na ortofotomapie. Dużo lepszym pomysłem jest wykorzystanie zdjęć lotniczych do detekcji takich pionowych obiektów. Na szczególną uwagę zasługują obrazy ukośne, ponieważ w przypadku takiego zbioru danych obiekt jest nie tylko poprawnie odwzorowany, ale także jest odwzorowany kilka razy z wielu kierunków. Umożliwia to wykrycie latarni na wielu obrazach, a zatem nawet jeśli obiekt jest niewidoczny na jednym obrazie, ponieważ jest na przykład częściowo zasłonięty przez drzewo, istnieje szansa, że zostanie odwzorowany na innym obrazie. Na przykład na kolejnym zdjęciu w szeregu lub na zdjęciach z sąsiednich szeregów.

Pamiętając o tym, że aby wytrenować model do detekcji obiektów, konieczne jest oznaczenie wielu tysięcy zdjęć, co jest czasochłonnym zadaniem. Etap pracy ręcznej (etykietowania zdjęć) jest często nieunikniony, jednak może być automatyzowany. Jedną z propozycji jest

rozwiązanie, które łączą różne dane fotogrametryczne jako dane wyjściowe w celu utworzenia zbioru danych do trenowania modelu. Zlokalizowanie obiektu (w odniesieniu do latarni) jako punktu na ortofotomapie może być wykonane znacznie szybciej niż etykietowanie każdego zdjęcia po kolei i rysowanie poligonów wokół obiektów. W następnym kroku można przyciąć fragment chmury do obiektu zainteresowania i przenieść współrzędne punktów chmury z układu terenowego do współrzędnych pikseli obrazu na podstawie elementów orientacji wewnętrznej i zewnętrznej. W odniesieniu do przeglądu i tego, jak latarnie odwzorowały się na różnego typu danych widać, że dane hybrydowe wykorzystane w pracy mogą być dobrym źródłem do takiego podejścia. Metodyka pozwala zatem uniknąć oznaczania każdego zdjęcia osobno i jednego obiektu kilka razy na dziesiątkach zdjęć. Podejście to zostało opracowane przez autorkę na wcześniejszym etapie badań i opisane w artykule (Zachar i in., 2022).

Spośród zebranych i przeanalizowanych typów obiektów małej architektury wybrano latarnie uliczne do dalszej części badawczej zawierającej wyniki eksperymentów. Latarnia uliczna to ważny obiekt z punktu widzenia urbanistyki i przestrzeni miejskiej, ale także interesujący pod względem detekcji ze względu na swoją charakterystykę - wysoki, smukły element. W porównaniu z innymi obiektami o bardziej regularnych kształtach latarnie wydają się być interesującym obiektem badawczym. Ponadto są to obiekty dość liczebne i występujące przeważnie na całym obszarze miasta. Ukośne zdjęcia lotnicze wydają się więc być dobrym źródłem danych do ekstrakcji informacji o takich obiektach w skali miejskiej.

Wybór dwóch różnych obiektów miejskich miał na celu zweryfikowanie potencjału wykorzystania różnych metod wykrywania obiektów na lotniczych zdjęciach ukośnych. Ponadto decyzja odnośnie wyselekcjonowania tych dwóch klas obiektów była podyktowana ich reprezentatywnością wśród innych klas malej architektury, a co za tym idzie możliwość implementacji metodyki do takich kategorii obiektów jak: banery reklamowe, balkony, znaki drogowe, słupy uliczne, słupy trakcyjne.

6. Analiza kompletności chmur punktów z gęstego dopasowania obrazów dla latarni ulicznych

Dobór źródła, na podstawie którego można wykrywać dane obiekty jest ściśle związany z tym czy dany obiekt jest widoczny i rozpoznawalny dla danego typu danych. Jak zostało wcześniej nakreślone odwzorowanie różnego typu obiektów, w szczególności gdy mowa o małych obiektach architektury miejskiej może znacznie się różnić na przykład pomiędzy chmurą punktów pozyskaną z gęstego dopasowania obrazów a danymi ALS.

Biorąc więc pod uwagę wykorzystanie różnych źródeł danych do wykrywania obiektów lub tworzenia zbiorów danych do szkolenia modeli, konieczne jest zweryfikowanie kompletności i dokładności tych danych. W pierwszym zaprojektowanym eksperymencie przeanalizowano chmury punktów z gęstego dopasowania zdjęć lotniczych w celu zbadania ich kompletności względem referencyjnych danych z lotniczego skanowania laserowego w odniesieniu do reprezentacji latarni.

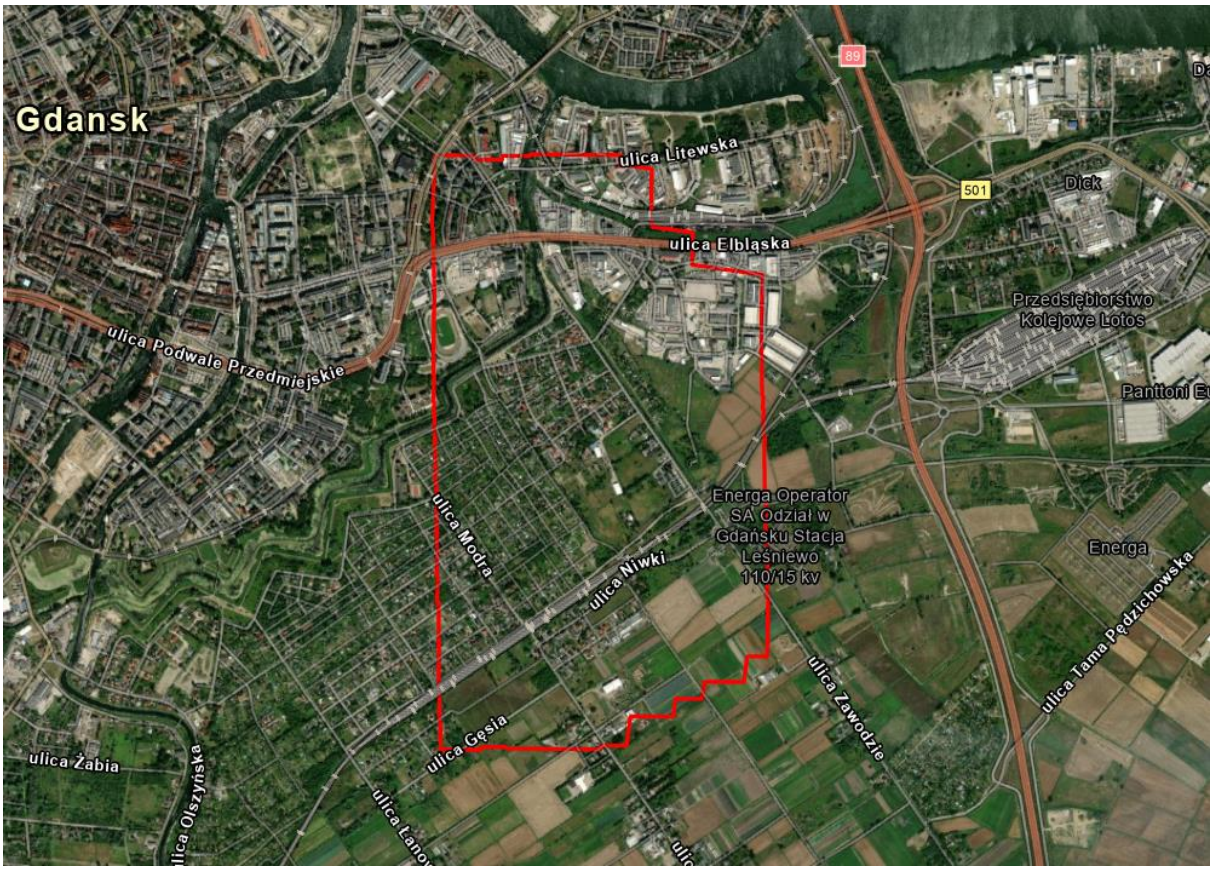
Pierwsza część badań dotyczy zatem dwóch aspektów. Z jednej strony wykonane eksperymenty miały na celu potwierdzenie założenia dotyczącego doboru zdjęć ukośnych jako najlepszego źródło danych do wykrywania latarni w przestrzeni miejskiej przy użyciu metod uczenia maszynowego, przy jednoczesnym pokazaniu problemów kompletności chmur punktów. Z drugiej zaś uwaga została poświęcona problemowi braku danych uczących i automatyzacji procesu etykietowania zdjęć. Wykorzystywanie różnych źródeł danych może być przydatne w kontekście żonglowania reprezentacjami i transferem informacji pomiędzy danymi fotogrametrycznymi. Dlatego też zaadaptowano opracowaną wcześniej metodykę (Zachar i in., 2022), w której wykorzystuje się fragmenty wyciętych chmur punktów dla obiektów zainteresowania jako dane wejściowe do tworzenia zbioru treningowego do detekcji na zdjęciach ukośnych. Współrzędne chmury punktów z układu terenowego przeliczone zostały do układu pikselowego zdjęcia i w ten sposób powstał zbiór uczący składający się ze zdjęć i informacji o ramkach ograniczających.

Metodologia ta umożliwiła zbadanie potencjału wykorzystania chmur punktów jako wsparcia przy tworzeniu szkoleniowego zbioru danych na ukośnych zdjęciach lotniczych. Przedstawiono zalety wykorzystywania różnego typu danych fotogrametrycznych oraz przechodzenia między reprezentacjami („żonglowania danymi”) w celu automatyzacji procesu tworzenia zbiorów uczących. Powstały zbiór uczący do detekcji latarni na zdjęciach ukośnych, który był wynikiem

pierwszego etapu badań posłużył do treningu sieci neuronowej YOLO w ramach drugiego eksperymentu.

6.1. Dane i obszar badań

Obszar badań (Rysunek 8) obejmował południowo-wschodnią część miasta Gdańsk o powierzchni około 2,5 km^2. Zbiór danych składał się z synchronicznie pozyskanych ukośnych zdjęć lotniczych charakteryzujących się wysokim pokryciem podłużnym i poprzecznym (ponad 80%) i chmur punktów LiDAR. Dane zostały pozyskane w maju 2022 r. za pomocą sensora Leica CityMapper-2. Terenowy wymiar piksela wynosi około 2,5 cm. Dane zostały przetworzone przy użyciu fotopunktów (GCP - ang. *Ground Control Point*), w tym dziesięciu fotopunktów (GCP) i czterech fotopunktów kontrolnych (ang. *check points*). Błąd reprojekcji wyniósł 0,585 piksela, a szczegółowe informacje o błędach GCP przedstawiono w Tabela 5.



Rysunek 8 Obszar opracowania (miasto Gdańsk) zaznaczony na czerwono.

Tabela 5 Podsumowanie uzyskanych dokładności na fotopunktach (GCP i Check points) dla bloku zdjęć.

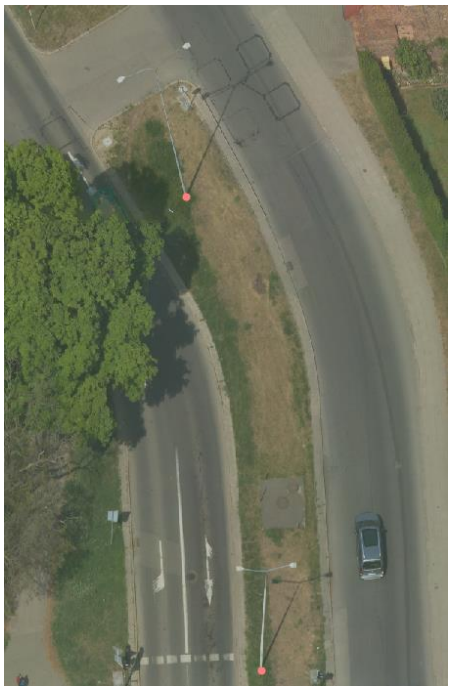
	Błąd RMS —X (cm)	Błąd RMS —Y (cm)	Błąd RMS —Z (cm)	RMSE (cm)
GCP RMSE	1.6	1.9	1.6	2.9
Check points RMSE	1.6	1.8	1.9	3.1

Poprawnie przetworzony blok zdjęć nadirowych i ukośnych został wykorzystany do wygenerowania gęstej chmury punktów i ortofotomapy dla obszaru opracowania. Oba te produkty fotogrametryczne, a także zdjęcia z poprawnie wyznaczonymi elementami orientacji zewnętrznej zostały wykorzystane podczas eksperymentów. Oprócz ukośnych zdjęć lotniczych wykorzystano również chmury punktów LiDAR, pozyskane synchronicznie wraz z wysokorozdzielczymi zdjęciami. Dane ALS dostarczone przez firmę zostały wcześniej przetworzone, więc nie było potrzeby wzajemnego wyrównania szeregów i ulepszania w stosunku do trajektorii. Średnia gęstość danych LiDAR to około 275 punktów/m^2. Chmury punktów z lotniczego skanowania laserowego w eksperymencie zostały wykorzystane jako dane referencyjne do weryfikacji kompletności chmur punktów z gęstego dopasowania zdjęć lotniczych.

6.2. Przygotowanie zbioru danych

Jak wspomniano wcześniej, głównym celem tego eksperymentu jest weryfikacja kompletności chmur punktów z gęstego dopasowania obrazu (DIM) pod kątem ich potencjału do wspierania automatyzacji tworzenia zbiorów danych uczących. Dokładna metodologia powstała we wcześniejszym etapie badań i została w pełni opisana i opublikowana w artykule autorki (Zachar i in., 2022). To samo podejście zastosowano w niniejszych eksperymentach.

Pierwszym krokiem było przygotowanie bazy danych z obiektami zainteresowania. Na potrzeby badania zdecydowano się na dwa typy latarni ulicznych - pojedyncze i podwójne (Rysunek 9). Lokalizacje latarni ulicznych zostały określone na podstawie ortofotomapy na poziomie gruntu. Utworzona została warstwa punktowa, która zawierała informacje o współrzędnych terenowych X, Y i Z i typie latarni.

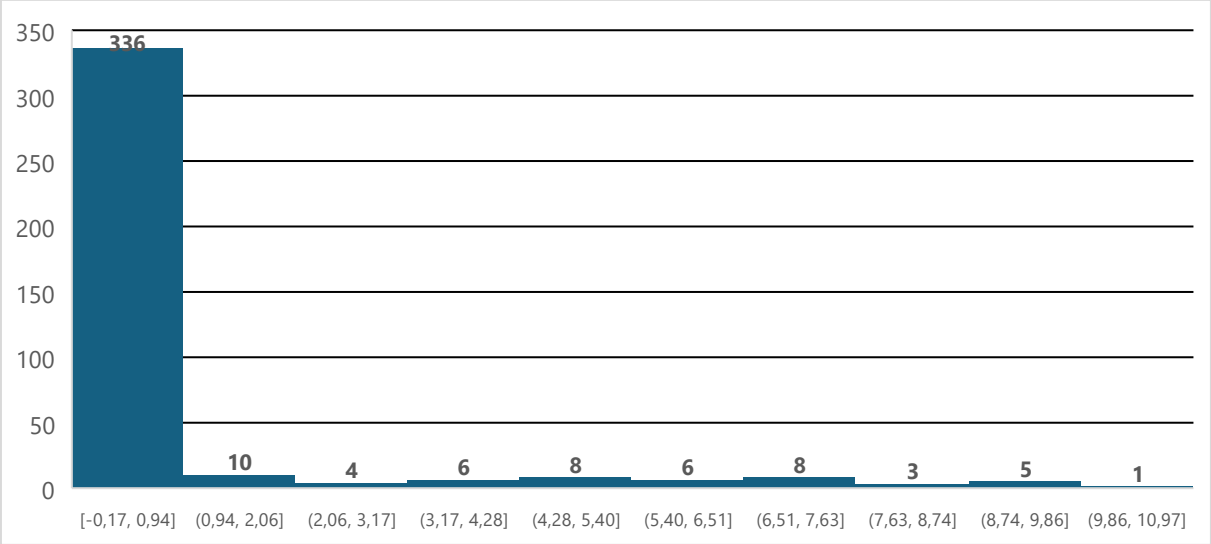
a)  b) 

Rysunek 9 Przykłady latarni ulicznych (a) - pojedynczych, b) podwójnych) zinwentaryzowanych na potrzeby eksperymentów.

Dla obszaru zainteresowania zaznaczono 409 pojedynczych latarni ulicznych. Następnie, przy użyciu tych punktów, przycięto chmury punktów ze skanowania laserowego (ALS) i gęstego dopasowania obraz. Aby mieć pewność, że wszystkie punkty chmury należące do obiektu zostaną uwzględnione w wyciętych fragmentach zdefiniowano warunek, że wszystkie punkty leżące w promieniu 2m od punktu wskazanego na gruncie należą do obiektu. Wycięte w ten sposób fragmenty zostały przeanalizowane i zweryfikowane pod kątem poprawności. Wszelkie niedoskonałości w bazie danych zostały skorygowane. W kilku przypadkach zdarzyło się, że punkt naziemny został przesunięty względem rzeczywistej pozycji obiektu, na co wpływ miały błędy wynikające z niedoskonałości wytworzonej ortofotomapy. Błędy te zostały wyeliminowane. Z całego zestawu latarni ulicznych 22 z nich zostały odrzucone z dalszej analizy. Były to głównie latarnie uliczne zlokalizowane wśród gęstych drzew lub krzewów, dla których możliwe było odwzorowanie tylko górnej części latarni ulicznej w chmurze punktów z dopasowania obrazu ze względu na brak penetracji roślinności dla tej metody.

Każda z 387 latarni ulicznych została przycięta tak, aby odfiltrowane zostały również punkty na gruncie i w rezultacie w wyciętych fragmentach znalazły się tylko punkty należące do obiektu bez żadnych szumów. Mając dwa zestawy danych (chmury punktów z ALS i chmury punktów DIM) możliwe było porównanie wysokości obiektów dla obu typów. Wszystkie

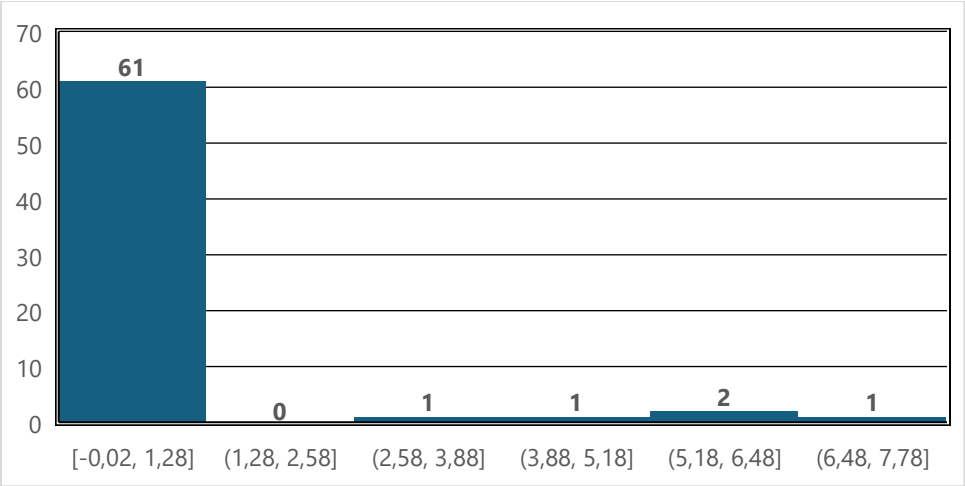
latarnie zostały porównane pod względem różnic wysokości między danymi referencyjnymi ALS a chmurami punktów z DIM.



Rysunek 10 Różnice wysokości między chmurami punktów ALS a chmurami uzyskanymi z DIM reprezentującymi pojedyncze latarnie uliczne.

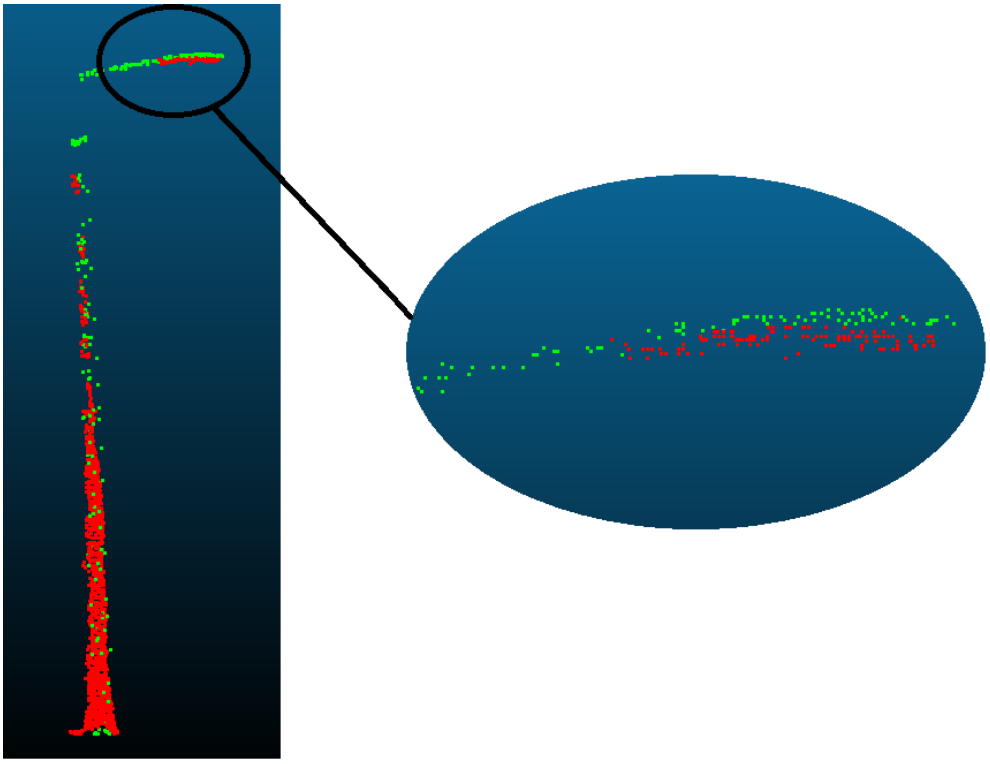
Wykres (Rysunek 10) przedstawia różnice wysokości latarni między chmurą punktów ALS a chmurą punktów z dopasowania zdjęć. Widać, że większość różnic mieści się w zakresie kilkunastu/kilkudziesięciu centymetrów. Pozostałe przypadki to te, dla których obiekt nie został odwzorowany w chmurze DIM lub odwzorowana została tylko jego część. Średnia różnica wysokości wynosi około 73 cm. Po odrzuceniu latarni, które w ogóle nie zostały odwzorowane w chmurach punktów z gęstego dopasowania obrazów, średnia ukształtowała się na poziomie 58 cm.

Zbiór danych obejmował również podwójne latarnie uliczne, których było 66 w obszarze zainteresowania. Podobnie jak w przypadku latarni pojedynczych każdy obiekt został sprawdzony pod kątem poprawności. Żaden obiekt nie został odrzucony na tym etapie. Weryfikacja różnic w wysokościach podobnie jak w przypadku pojedynczych latarni ulicznych wykazała, że większość wartości mieści się w zakresie do około 1 m (Rysunek 11). Średnia dla latarni podwójnych wyniosła 46 cm.



Rysunek 11 Różnice wysokości między chmurami punktów ALS a chmurami uzyskanymi z DIM reprezentującymi podwójne latarnie uliczne.

Podczas analizy wizualnej i sprawdzania bazy danych obiektów i wyciętych fragmentów chmur punktów zauważono, że dla większości obiektów chmury punktów ALS wykazywały wyższe wartości wysokości (Rysunek 12). Podjęto próbę znalezienia przyczyny i zweryfikowania, skąd mogą pochodzić te różnice. Przeanalizowano odległości między chmurami punktów na innych obiektach, takich jak droga o twardej nawierzchni. Nie znaleziono żadnych przesunięć statystycznych i stwierdzono, że błędy te wynikają z różnic w metodach pozyskiwania obu produktów.

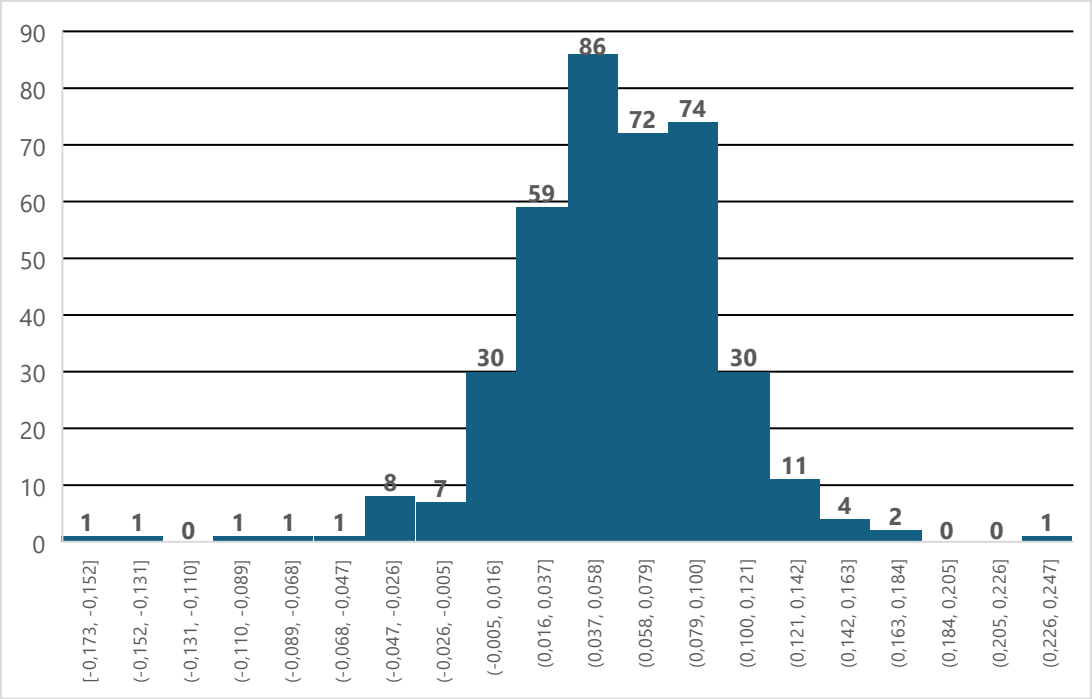


Rysunek 12 Przykład różnic wysokości między chmurą z gęstego dopasowania obrazów (kolor czerwony) a chmurą z lotniczego skaningu laserowego (kolor zielony) dla pojedynczej latarni ulicznej.

Aby baza danych była w pełni poprawnie przygotowana, zweryfikowano również, na ilu zdjęciach jest odwzorowana każda z latarni. Ponieważ obszar zainteresowania został wybrany tak, aby obejmował część bloku o pełnym pokryciu zdjęciami, każdy obiekt w tym obszarze powinien znajdować się na tej samej liczbie zdjęć. Sprawdzono, czy latarnie, które nie zostały odwzorowane na chmurach DIM nie znajdowały się na przykład na krawędziach obszaru i czy aby na pewno liczba zdjęć, na których obiekt został odwzorowany nie była mniejsza względem oczekiwanej. Nie znaleziono żadnego wzorca, który wskazywałby na wpływ pokrycia. Zdarzało się również tak, że latarnie znajdujące się w środku nie odwzorowały się na chmurze punktów DIM pomimo pełnego pokrycia zdjęciami z wielu kierunków.

W kolejnym kroku przygotowania danych wejściowych do eksperymentów odrzucono jeszcze ze zbioru osiem latarni pojedynczych, które pomimo dobrego odwzorowania na danych ALS, w ogóle nie zostały odwzorowane na chmurach DIM. W rezultacie do dalszej analizy statystycznej wprowadzono 379 pojedynczych latarni i 66 podwójnych latarni (razem 445 obiektów).

W celu bardziej przejrzystego przedstawienia wyników dla różnic wysokości między chmurami punktów sporządzono wykres rozkładu wartości dla latarni o najmniejszych różnicach. Z zestawu wybrano tylko te latarnie, dla których różnice były mniejsze niż 25cm. Spośród 445, 389 obiektów mieściło się w tym zakresie.



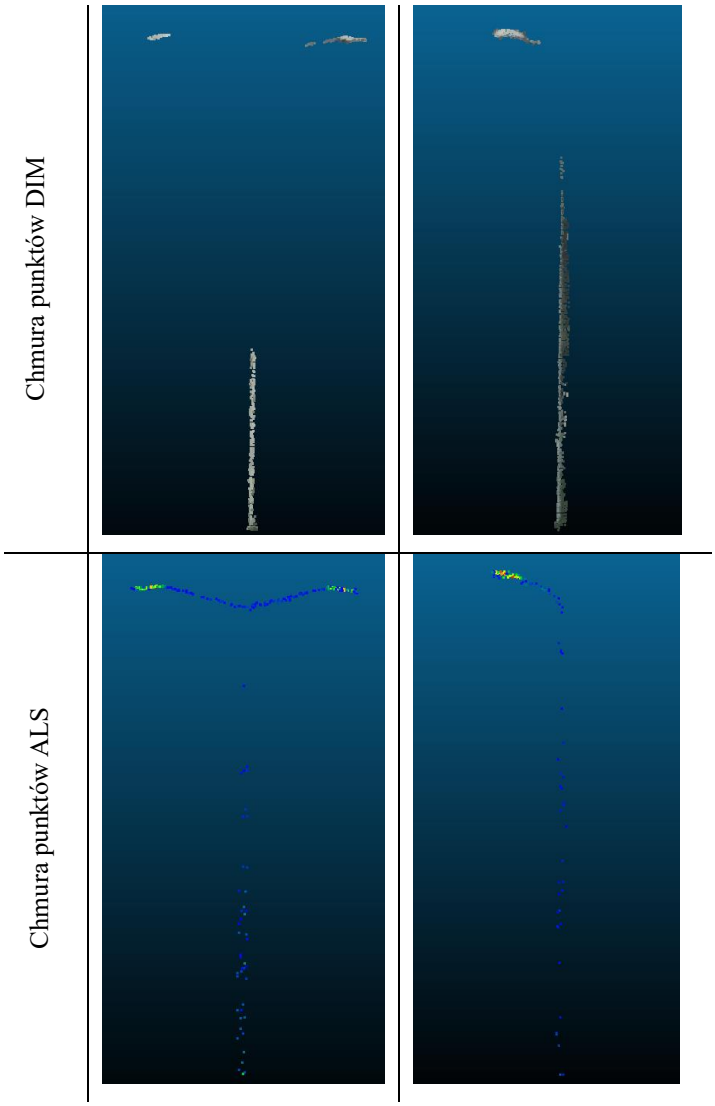
Rysunek 13 Podsumowanie różnic wysokości między chmurami ALS i DIM dla latarni po odrzuceniu najbardziej odstających wartości.

Różnice między wysokościami latarni ALS i DIM przedstawiono na histogramie powyżej (Rysunek 13). Widać, że większość z nich to różnice rzędu kilku centymetrów. Większość obiektów mieści się w zakresie 3,7-5,8 cm. Jak wspomniano wcześniej, dokładna ocena wizualna chmur punktów i analiza tych latarni, dla których różnice były powyżej 25 cm, nie pozwoliła na znalezienie odpowiedzi, dlaczego niektóre zostały odwzorowane w chmurze punktów, a inne nie.

Poniżej znajduje się przykład latarni tego samego typu o podobnej wysokości (Rysunek 14). Latarnia uliczna po lewej stronie jest odwzorowana w obu zestawach (DIM i chmura punktów ALS), podczas gdy latarnia uliczna po prawej stronie nie jest w pełni odwzorowana w DIM. W rzeczywistości tylko dolna część blisko ziemi znajduje się w chmurze punktów. Oba obiekty zostały odwzorowane na tej samej liczbie obrazów.

Rysunek 14 Przykład poprawnie odwzorowanej latarni w chmurze na podstawie dopasowania obrazu (latarnia po lewej stronie) oraz latarni nieodwzorowanej w chmurze punktów z dopasowania (latarnia po prawej stronie).

Ostateczny zbiór danych po wszystkich weryfikacjach składał się z 379 pojedynczych latarni i 66 podwójnych latarni. Przykłady obiektów wyciętych z chmury punktów pokazano na poniższym rysunku (Rysunek 15).



Rysunek 15 Przykłady latarni ulicznych (wyciętych z chmury punktów) z docelowej bazy danych: na górze dla chmury z dopasowania obrazów, na dole dla chmury LiDAR.

6.3. Kompletność chmur punktów

Podejście, które wykorzystuje chmury punktów DIM jako źródło do tworzenia zbioru danych uczących dla metod głębokiego uczenia się na obrazach ukośnych, powinno zakładać, że każdy obiekt jest poprawnie reprezentowany i w pełni odwzorowany w chmurze punktów. Stąd też wynikała wcześniejsza wnikliwa analiza, która pozwoliła na wyeliminowanie niektórych obiektów.

Aby ramka ograniczająca obiekt zainteresowania na zdjęciu lotniczym, która została utworzona przez reprojekcję współrzędnych punktów terenowych do współrzędnych w układzie pikselowym, była poprawna, chmura punktów powinna poprawnie reprezentować obiekt. Oznacza to, że zarówno górna, jak i dolna część latarni ulicznej powinna zostać odwzorowana.

Jednak, jak wspomniano wcześniej, zdarza się, że niektóre elementy nie są widoczne na danym typie danych.

Postanowiono więc przeprowadzić eksperyment, podczas którego zbadana zostanie kompletność chmur punktów z dopasowania obrazów dla latarni. Jako dane referencyjne wykorzystano dane LiDAR.

Kompletność chmur punktów można zbadać za pomocą rozkładu punktów w chmurze. Znając wysokość obiektu, można ocenić każdy wycięty fragment chmury punktów (przedstawiający latarnię), porównując zmiany percentyli wysokości punktów należących do obiektu. Interpretacja rozkładu percentyli w chmurze punktów może być wykorzystywana w różnych zastosowaniach, na przykład w analizie roślinności. Istnieją publikacje, w których przedstawiono wyniki badania kompletności chmur punktów do tego typu zadań. Autorzy (White i in., 2015) w swojej pracy wykorzystali metryki dla danych ALS i SGM (*ang. Semi-Global Matching*) w celu zbadania, w jaki sposób te dwa źródła danych charakteryzują pionową strukturę lasu.

Większość zastosowań badania rozkładu gęstości chmury punktów odnosi się do zastosowań leśnych (Næsset, 2002; Weinmann i in., 2017; Moudrý i in., 2023). Cechy geometryczne chmur punktów mogą być przydatne na przykład do rozpoznawania podczas klasyfikacji gatunków drzew (Michałowska i Rapiński, 2021). Zarówno statystyki wysokości, jak i cechy opisujące rozkład punktów, takie jak kurtoza, skośność, penetracja, percentyle rozkładu wysokości punktów w klastrze, gęstość i liczba punktów w 10-poziomowej warstwie wysokości drzewa, pozwalają na ekstrakcję cech dla poszczególnych obiektów w źródłowej chmurze punktów. Można podać jeszcze inne zastosowania. W artykule (Yu i in., 2017) opisane zostały badania, których celem była ocena potencjału wielospektralnych danych lotniczego skanowania laserowego do wykrywania pojedynczych drzew i klasyfikacji gatunków drzew. Podobne analizy dla tego typu danych zostały przedstawione w pracach Amiri i in., 2018 oraz Axelsson i in., 2018.

Powyżej przytoczone badania koncentrują się głównie na percentylach rozkładu wysokości. Takie podejście do badania pionowego rozkładu punktów w chmurze może być przydatne w wielu zastosowaniach, zarówno weryfikując kompletność mapowania danych elementów, jak i wyszukując obiekty o określonych cechach. Innym przykładem, w którym zaproponowano algorytm bazujący na percentylach wysokości było rozpoznawanie słupów drogowych (Pu i in., 2011). Rozwój wspomnianej metody został opisany w pracy magisterskiej (Li, 2013).

Podsumowując, analiza pionowego rozkładu chmur punktów przy użyciu różnych metryk, takich jak percentyle, jest skuteczną strategią w różnych obszarach zastosowań. Dlatego też po przeanalizowaniu literatury i porównaniu różnych metod badania kompletności chmur punktów zdecydowano się na podejście wykorzystujące pomiar pionowego rozkładu gęstości punktów w chmurze. Metodyka została zaadaptowana na potrzeby eksperymentów, a wyniki przedstawiono i omówiono poniżej.

Percentyle wysokości zostały przeliczone dla każdej chmury punktów przedstawiającej latarnie, wskazując wysokość, poniżej której zarejestrowano określony procent punktów. Jeśli chodzi o interpretację percentyli w kontekście chmury punktów, jak zostało wspomniane, dostarcza ona informacji o pionowym rozkładzie punktów na różnych wysokościach. Rozkłady gęstości odzwierciedlają proporcję punktów w określonym zakresie wysokości do całkowitej liczby punktów. Choć dolny kwantyl (25%), górny kwantyl (75%) oraz mediana (50. percentyl) dostarczają danych statystycznych dla zbioru, postanowiono dokładniej zbadać obserwacje, aby umożliwić lepsze zrozumienie analizowanego zjawiska. Wyniki zostały podzielone na 20 przedziałów, od 5% do 95% w 5% przyrostach.

Do analizy statystyk kwantylowych wybrano tylko te latarnie, dla których różnica w wysokości między chmurą ALS i DIM nie przekraczała wartości 25cm, resztę obiektów odrzucono. Ostatecznie rozkład kwantyli obliczono dla 327 pojedynczych latarni. Kwantyle były obliczane w kolejnych krokach. Dla każdego obiektu wartości współrzędnej „Z" każdego punktu w chmurze została znormalizowana poprzez podzielenie przez wysokość z danych ALS, stanowiących dane referencyjne w badaniach.

W celu lepszego zrozumienia wyników poddano interpretacji wartości dla kilku obiektów z zestawu. Poniżej zamieszczono przykładową tabelę z wynikami dla jednej z latarni, która pokazuje współczynnik kwantyla i odpowiadającą mu wartość kwantyla (% wysokości) dla danego zestawu danych (Tabela 6). Współczynnik kwantylowy reprezentuje odsetek danych poniżej określonej wartości. Waha się od 0,0 do 1,0, gdzie 0,0 reprezentuje wartość minimalną, a 1,0 reprezentuje wartość maksymalną. Wartość kwantyla reprezentuje wartość danych, poniżej której znajduje się określony procent danych. Na przykład wartość kwantyla 0,1 oznacza, że 10% danych znajduje się poniżej tej wartości. Ogólnie rzecz biorąc, wyniki te dostarczają informacji o rozkładzie zbioru danych. Pomagają zidentyfikować konkretne wartości, które dzielą dane na określone proporcje.

Tabela 6 Kwantyle wysokości dla jednego przykładu pojedynczej latarni.

Quantile Factor	Quantile Value
0.00	0.000
0.05	0.005
0.10	0.009
0.15	0.014
0.20	0.020
0.25	0.032
0.30	0.044
0.35	0.064
0.40	0.094
0.45	0.121
0.50	0.143
0.55	0.168
0.60	0.198
0.65	0.229
0.70	0.266
0.75	0.331
0.80	0.627
0.85	0.638
0.90	0.651
0.95	0.988

W przytoczonym przykładzie można zauważyć, że dla kwantyla 0,50 (50. percentyl) wartość wynosi 0,143. Oznacza to, że 50% danych jest mniejsze lub równe 0,143, co stanowi mniej niż 15% wysokości latarni. Ponadto dla niektórych kwantyli wartości są małe (np. 0,005 dla kwantyla 0,05), co sugeruje, że większość danych koncentruje się wokół niższych wartości. Inne przykłady wyników rozkładu punktowego dla latarni przedstawiono w Tabela 7. W pierwszym wierszu widać fragment z chmury punktów DIM, w którym brakuje dolnej części obiektu. Drugi wiersz pokazuje ten sam obiekt dla chmury punktów ALS, gdzie rozkład punktów jest bardziej ciągły na całej wysokości.

Tabela 7 Przykładowe wyniki dla tego samego obiektu pokazujące pionowy rozkład gęstości chmury punktów: (a) chmura punktów DIM, (b) chmura punktów ALS.

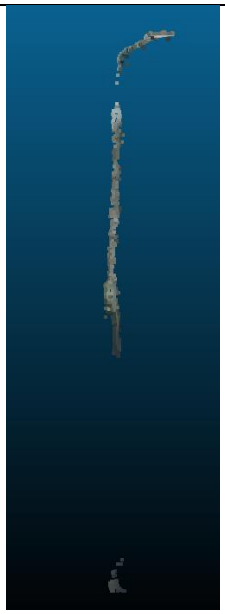
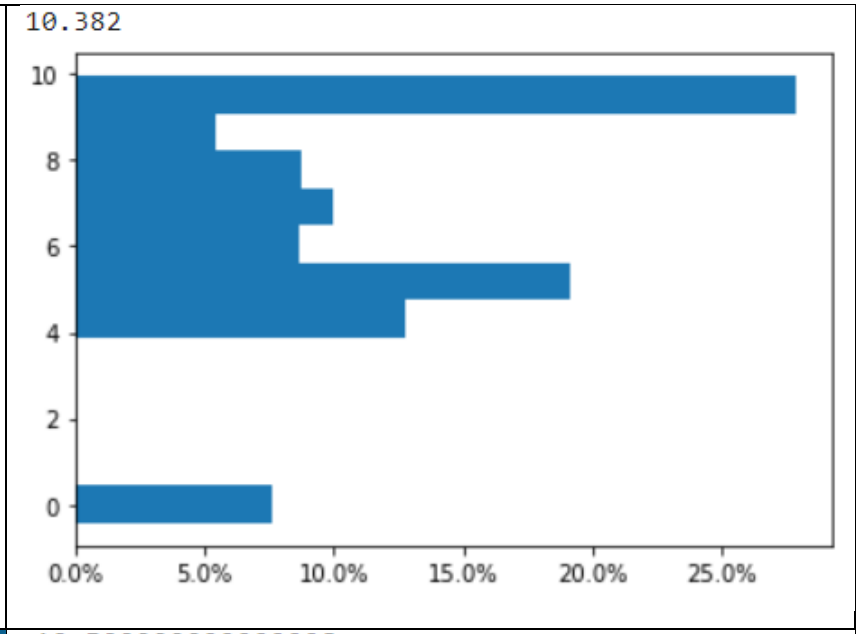
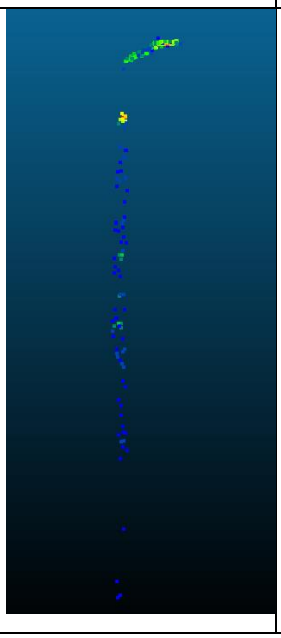
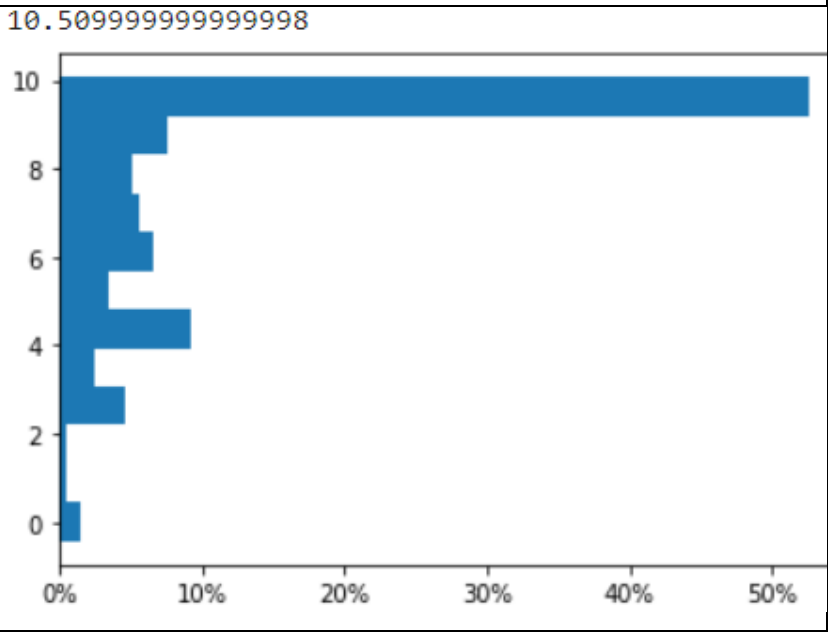
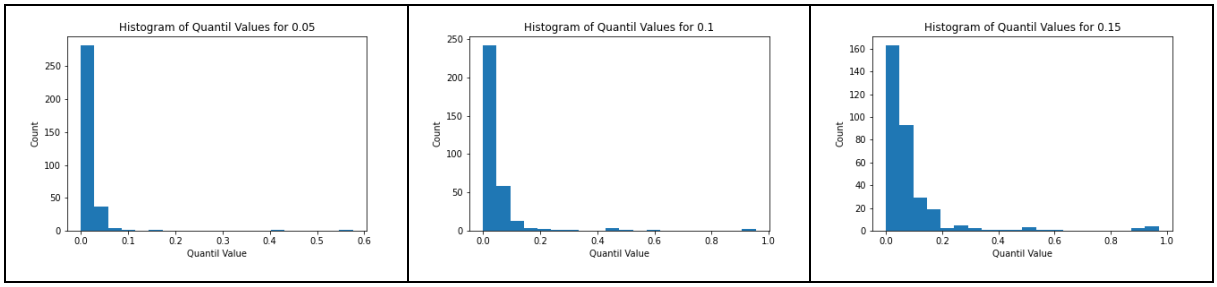
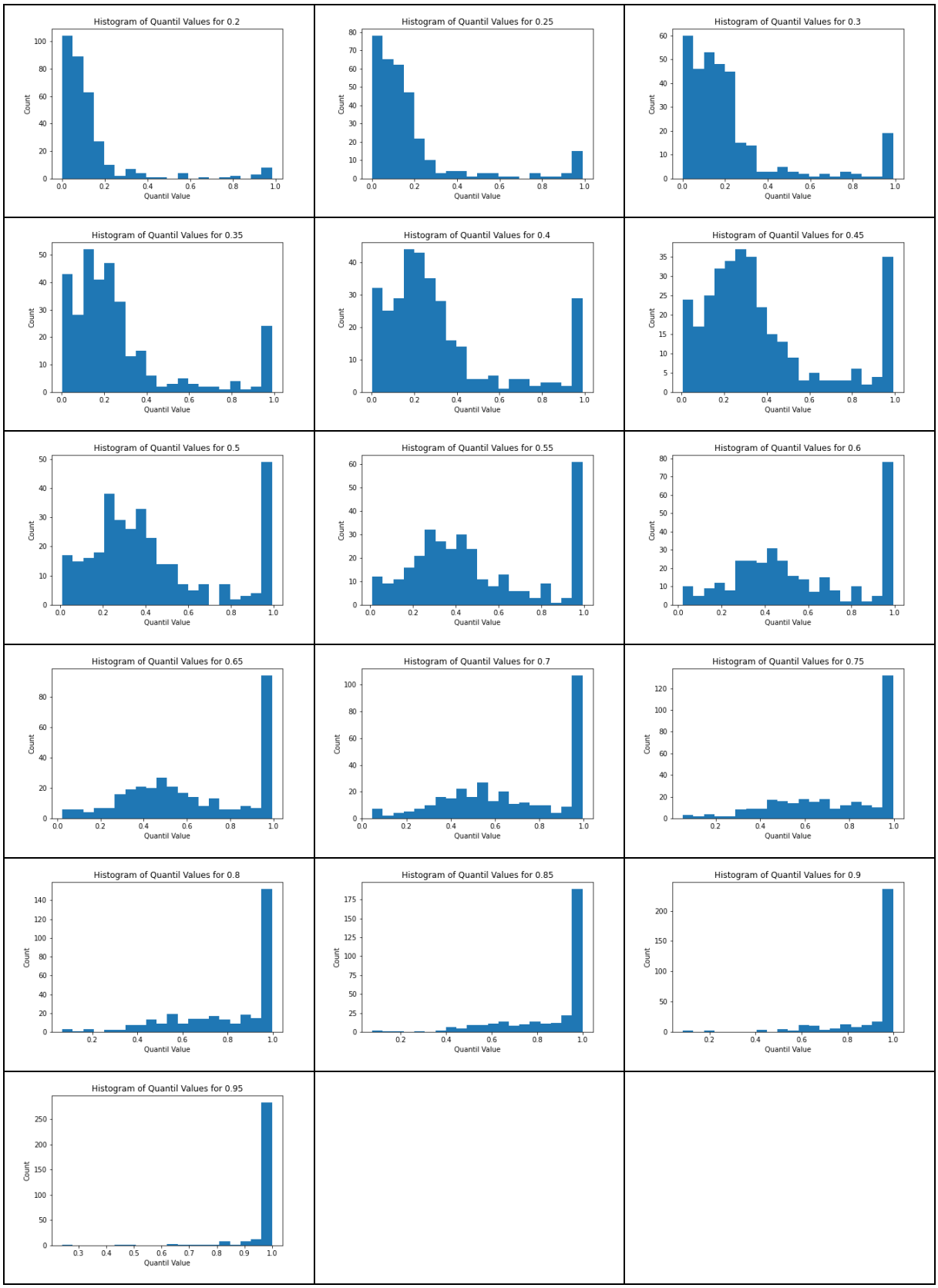
a)		
b)		

Tabela 8 Histogramy dla pojedynczych latarni dla wszystkich współczynników kwantylowych, gdzie wartości oznaczają znormalizowane wysokości obiektów.





Na podstawie powyższych histogramów można stwierdzić, że największa gęstość punktów dla chmur punktów z dopasowania obrazu znajduje się na dole latarni (tj. na poziomie gruntu) i w górnej części obiektu. Podczas analizy histogramów skupiono się na dolnym kwantylu

(25%), drugim kwantylu (50%), który jest odpowiednikiem mediany, oraz górnym kwantylu (75%). W przypadku kwantyla 25% wartości skupiają się między 0,0 a 0,3 wysokości latarni. Oznacza to, że 25% punktów dla większości latarni znajduje się w dolnej części obiektu. Niewielki wzrost można również zaobserwować w górnej części niektórych obiektów. Współczynnik kwantyla 50% dla większości latarni również znajdował się poniżej połowy wysokości obiektów. Jednak znacznie większy pik można zaobserwować dla wartości powyżej 0,95 wysokości. Analizując górny kwantyl (75%), dla około 1/3 obiektów 75% punktów znajduje się w górnej części. Wyniki te są zgodne ze wstępną oceną wizualną. Spośród 66 podwójnych latarni odrzucono 6, dla których różnice wysokości były większe niż 25 cm. Wszystkie kolejne kroki zostały wykonane analogicznie. Wysokości zostały znormalizowane i dla takich danych policzono statystyki. Przykład, dla którego wyniki przedstawiono w Tabela 9, różni się od rozkładu przedstawionego dla pojedynczej latarni. Widać, że tylko w niewielkim stopniu został odwzorowany w chmurze punktów, co można również zauważyć interpretując tabelę wartości kwantylowych.

Tabela 9 Kwantyle wysokości dla jednej przykładowej latarni (chmura DIM).

Quantile Factor	Quantile Value
0.00	0.000
0.05	0.955
0.10	0.975
0.15	0.980
0.20	0.981
0.25	0.981
0.30	0.982
0.35	0.983
0.40	0.985
0.45	0.986
0.50	0.986
0.55	0.986
0.60	0.987
0.65	0.987
0.70	0.988
0.75	0.989
0.80	0.989
0.85	0.990
0.90	0.991
0.95	0.993

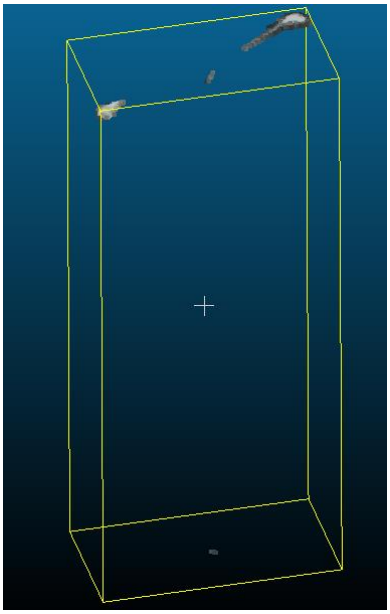
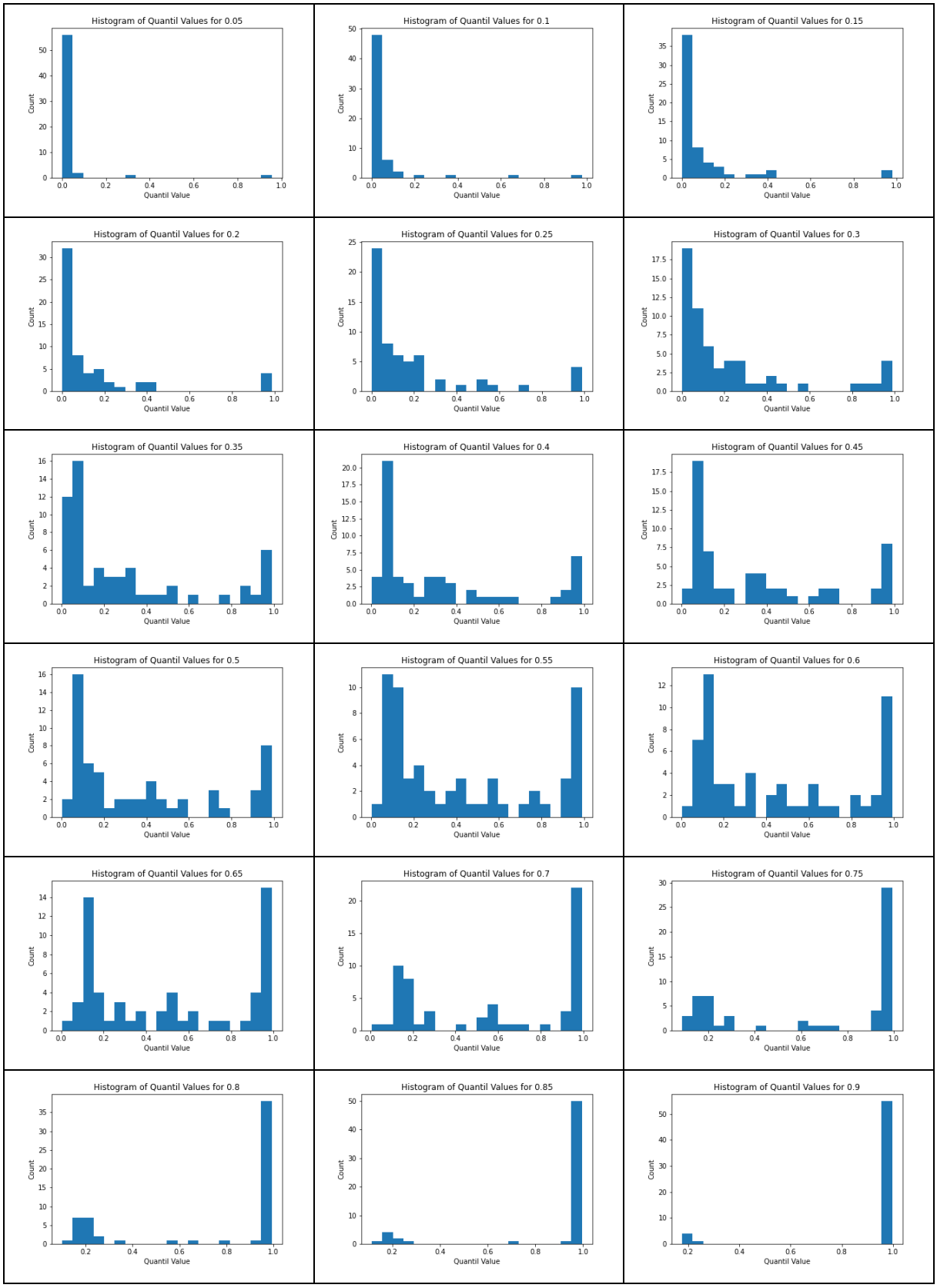
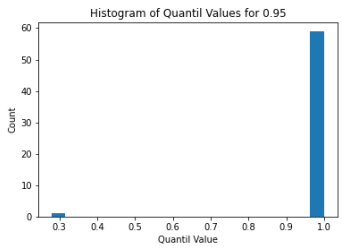


Tabela 10 Histogramy dla podwójnych latarni dla wszystkich współczynników kwantylowych, gdzie wartości oznaczają znormalizowane wysokości obiektów.



Wyzwanie związane z interpretacją wielu histogramów było motywacją do kompresji wyników. Licząc kwantyle od 0 do 95% dla każdej latarni, wynikiem jest ponownie obliczona wartość wysokości dla każdej z nich, która mieści się w zakresie 0-1, ponieważ są one znormalizowane do referencyjnych danych LiDAR. Następnie, dla każdego kwantyla po kolei, wartości są wybierane, a następnie dzielone na 20 równych przedziałów (tak jak w przypadku kwantyli) i liczone, ile latarni ulicznych znajdowało się w tym przedziale. Takie podejście umożliwiło stworzenie histogramów 3D. Gdzie na osiach poziomych znajdują się kwantyle i ich wartości, a oś pionowa wskazuje liczbę latarni ulicznych (Tabela 11, Tabela 12).

Tabela 11 Histogramy 3D dla pojedynczych i podwójnych latarni dla chmur punktów z gęstego dopasowanie obrazów. Osie poziome to kwantyle i ich wartości (znormalizowane wysokości), a oś pionowa wskazuje liczbę obiektów – latarni znajdujących się w danym zakresie.

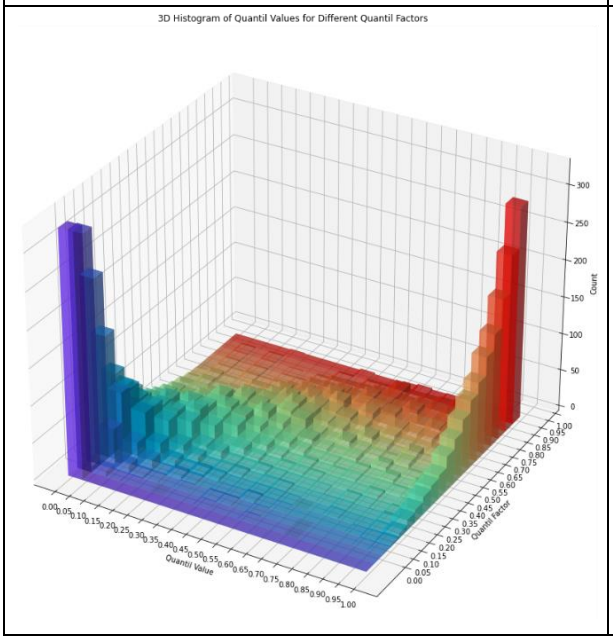
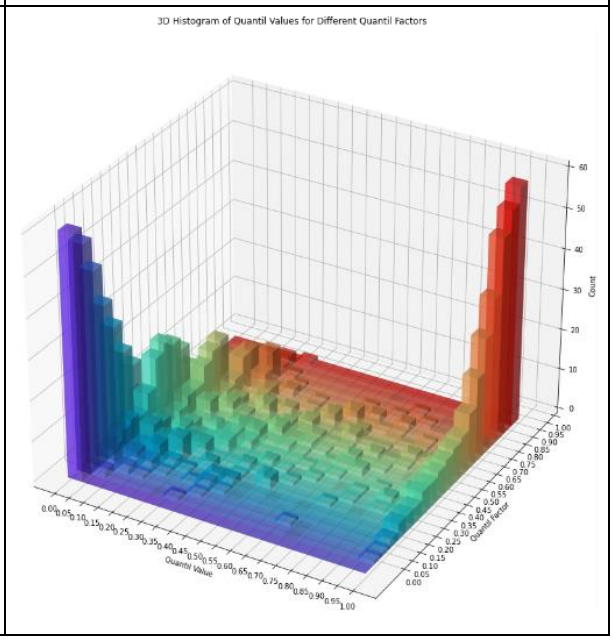
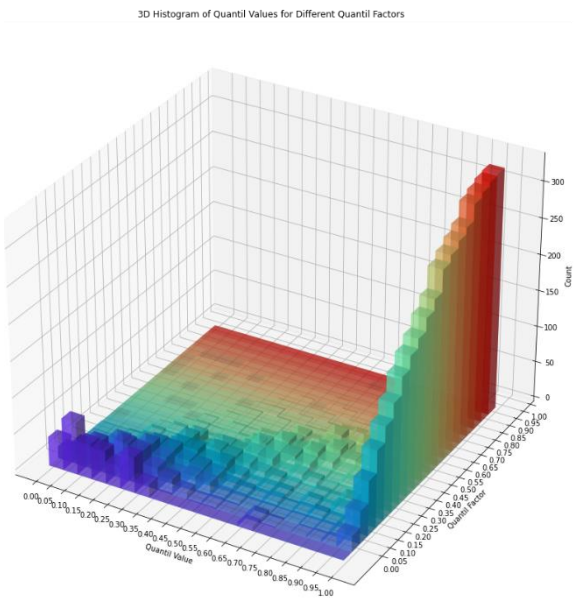
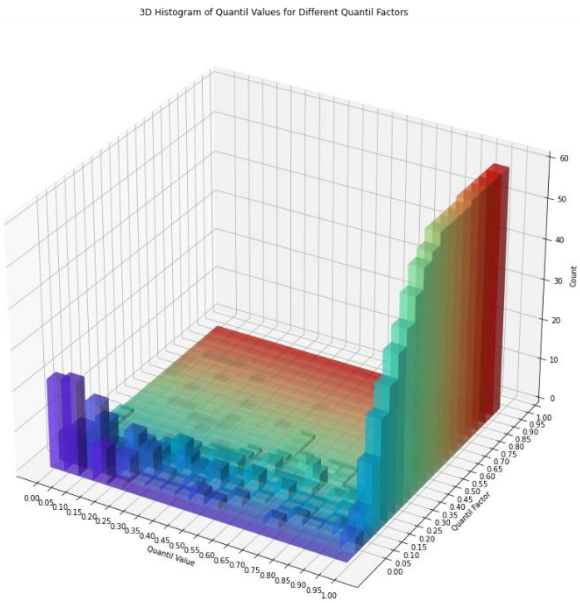
LATARNIE POJEDYNCZE	LATARNIE PODWÓJNE
	

Tabela 12 Histogramy 3D dla pojedynczych i podwójnych latarni dla chmur punktów z referencyjnego ALS. Osie poziome to kwantyle i ich wartości (znormalizowane wysokości), a oś pionowa wskazuje liczbę obiektów – latarni znajdujących się w danym zakresie.

LATARNIE POJEDYNCZE	LATARNIE PODWÓJNE
	

Wyniki wykonanych eksperymentów potwierdzają to, że występują problemy związane z kompletnością chmur punktów z dopasowania zdjęć lotniczych. Najtrudniejsze oczywiście są przypadki, w których latarnia nie została w ogóle odwzorowana w chmurze punktów. Biorąc jednak pod uwagę, że aby chmury punktów mogły być wykorzystane jako wsparcie przy tworzeniu zestawu uczącego do detekcji latarni na ukośnych zdjęciach lotniczych, dane muszą być poprawne i kompletne. Częściowe odwzorowanie obiektu (na przykład tylko górna część obiektu) w chmurze punktów może również stanowić problem.

Powyższe wyniki pokazały, że kompletność chmur punktów z dopasowywania obrazów jest ważną kwestią przy wykorzystywaniu tego typu danych do tworzenia zbioru danych szkoleniowych przy użyciu metod półautomatycznych. Jest to szczególnie ważne w przypadku obiektów takich jak latarnie uliczne, które są wysokimi i smukłymi obiektami.

Powyższe badania miały na celu zbadanie kompletności chmur punktów z dopasowania obrazów i zweryfikowanie potencjału wykorzystania chmur DIM w celu automatyzacji procesu tworzenia zbioru uczącego. Oprócz badań związanych z analizą rozkładu punktów w chmurze dla latarni zweryfikowano poprawność utworzonego w automatyczny sposób zbioru danych składającego się ze zdjęć i ramek ograniczających obiekty. W tym celu współrzędne punktów chmury zostały przeliczone z układu terenowego na współrzędne pikselowe, co umożliwiło

określenie ramek ograniczających wyrażonych w pikselach dla każdego obiektu. W ten sposób utworzono spójny zbiór danych, który może być zarówno zbiorem testowym, jak i treningowym dla modeli głębokiego uczenia maszynowego.

Dokonano kolejnej analizy polegającej na porównaniu ramek ograniczających wygenerowanych na podstawie chmur punktów z gęstego dopasowania obrazów względem referencyjnych ramek z danych ALS. Wykorzystano do tego metrykę Intersection over Union (IoU). Poniżej (Tabela 13) przedstawiono porównanie ramek ograniczających powstałych na podstawie dwóch różnych źródeł danych. Jako dane referencyjne posłużyły ramki ograniczające uzyskane na podstawie chmur punktów z lotniczego skanowania laserowego, a te pozyskane na podstawie chmur punktów z gęstego dopasowania obrazów były poddane ocenie. W (Tabela 14) przedstawiono wyniki po odrzuceniu 12 latarni, dla których odwzorował się tylko kawałek na dole (poniżej 1m).

Tabela 13 Wyniki porównania ramek ograniczających uzyskanych na podstawie chmur punktów z gęstego dopasowania obrazów względem referencyjnych ramek ograniczających na podstawie danych ALS. Jako metrykę oceny zastosowano współczynnik IoU (Intersection over Union), wyrażony w %.

Kierunek	Liczba latarni	IoU [%]
Nadir (pionowy)	6433	69,38
Forward (w przód)	9791	75,90
Backward (wstecz)	10251	76,15
Left (w lewo)	7808	75,70
Right (w prawo)	9359	75,64

Tabela 14 Wyniki porównania ramek ograniczających uzyskanych na podstawie chmur punktów z gęstego dopasowania obrazów względem referencyjnych ramek ograniczających na podstawie danych ALS. Jako metrykę oceny zastosowano współczynnik IoU (Intersection over Union), wyrażony w %. Wynik ten uzyskano po odrzuceniu 12 latarni, dla których tylko kawałek latarni na dole się odwzorował (poniżej 1m).

Kierunek	Liczba latarni	IoU [%]
Nadir (pionowy)	6076	71,23
Forward (w przód)	9264	77,88
Backward (wstecz)	9686	78,27
Left (w lewo)	7353	77,62
Right (w prawo)	8910	77,59

Pełne odwzorowanie obiektów w chmurach punktów z lotniczego skanowania laserowego umożliwiło utworzenie kompletnego zestawu referencyjnego do wykrywania na ukośnych zdjęciach lotniczych.

7. Detekcja latarni z wykorzystaniem modelu YOLO

Utworzony zestaw danych, który był wynikiem pierwszej części badań składał się z ukośnych i pionowych zdjęć lotniczych oraz z plików zawierających informacje o ramkach ograniczających dla latarni. Ramki te powstały z przeniesienia wyciętych fragmentów chmur punktów na zdjęcia, aby pominąć etap związany z manualnym tworzeniem zbioru uczącego. Zbiór ten w dalszej części pracy został wykorzystany jak zestaw do trenowania sieci YOLOv8. Celem eksperymentu była analiza wykonalności zadania detekcji z wykorzystaniem wytrenowanego na podstawie powstałego w taki automatyczny sposób zbioru treningowego.

Wytrenowanie modelu YOLOv8 zostało poprzedzone przygotowaniem danych. Jeśli chodzi o trening modelu konieczne było podzielenie wysokorozdzielczych zdjęć na kafelki o rozmiarze 640x640 pikseli. Ponadto pliki zawierające informacje o ramkach ograniczających zostały przekonwertowane do odpowiedniego formatu, a położenie ramki zostało zamienione odpowiednio w odniesieniu do szerokości i wysokości kafelka. Poniżej (Tabela 15) zestawienie liczby zdjęć z każdego kierunku, które uwzględnione w procesie treningu oraz walidacji.

Tabela 15 Zestawienie liczby kafelków w zbiorze treningowym, walidacyjnym i testowym wykorzystanych do treningu modelu do detekcji latarni na zdjęciach lotniczych.

Kierunek	Liczba zdjęć w zestawie treningowym	Liczba zdjęć w zestawie walidacyjnym	Liczba zdjęć w zestawie testowym
Nadir	2727	1168	2000
Left	3442	1476	2000
Right	4461	1913	2000
Backward	5000	2142	2000
Forward	4601	1973	2000
RAZEM	**20231**	**8672**	**10000**

Przed przystąpieniem do uczenia sieci spośród całego zbioru 38903 kafelków 10000 z nich pozostawiono do wykonania niezależnej kontroli (po dwa tysiące dla każdego z kierunków). Postanowiono, że na potrzeby treningu pozostały zestaw zostanie podzielony odpowiednio 70% zdjęć w zestawie uczącym i 30% zostanie wykorzystanie do walidacji podczas treningu. W efekcie model został wytrenowany na podstawie 28903 kafelków, z czego 20231 było danymi treningowymi, a 8672 walidacyjnymi.

Dane walidacyjne były wykorzystane podczas procesu uczenia w celu dostrajania parametrów (hiperparametrów) modelu, umożliwiając sposób oceny wydajności modelu podczas szkolenia. Dzięki takiemu podejściu możliwa była ocena wydajności modelu, dostosowanie treningu w celu uniknięcia nadmiernego lub niedostatecznego dopasowania. Dane testowe zaś zostały wykorzystane po wytrenowaniu modelu w celu przeprowadzenia oceny, jak model poradzi sobie z danymi, których „nie widział" podczas treningu, co umożliwiło oszacowanie zdolności modelu do uogólniania na nowe dane.

Model YOLOv8 został stworzony w PyTorch i działa zarówno na CPU, jak i GPU. Proces uczenia został przeprowadzony na maszynie wyposażonej w NVIDIA GeForce RTX 3080Ti oraz 128GB pamięci RAM. Jeśli chodzi o szczegóły środowiska to wykorzystano Python w wersji 3.12 oraz technologię CUDA w wersji 12.4. Wytrenowany model w efekcie składa się z 295 warstw i 25856899 parametrów. Podstawowe parametry uczenia zestawiono poniżej (Tabela 16).

Tabela 16 Parametry dostosowane do treningu YOLOv8 do detekcji latarni na zdjęciach ukośnych.

Liczba epok	100
Batch size (oznacza liczbę obrazów przetwarzanych przed kolejną aktualizacją wewnętrznych parametrów modelu)	16
Image size (rozmiar kafelka)	640x640
Patience (liczba epok, po upływie których należy odczekać bez poprawy wskaźników walidacji przed wcześniejszym zatrzymaniem treningu)	50

Dokładność (Precision - P) oraz czułość (Recall - R) to podstawowe metryki oceny modeli głębokiego uczenia, wyrażone odpowiednio wzorami (1), (2):

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (1)$$

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (2)$$

TP (*true positives*) – predykcja pozytywna, zgodne ze stanem faktycznym;

FP (*false positives*) – predykacja pozytywna, faktycznie zaobserwowana klasa negatywna;

FN (*false negatives*) – predykacja negatywna, faktycznie zaobserwowana klasa pozytywna.

Dokładność (*precision*) określa odsetek poprawnych wyników pozytywnych wśród wszystkich pozytywnych predykcji, oceniając zdolność modelu do unikania fałszywych wyników pozytywnych. Czułość (*recall*) oblicza odsetek wyników prawdziwie pozytywnych wśród wszystkich rzeczywistych wyników pozytywnych, mierząc zdolność modelu do wykrywania wszystkich instancji danej klasy.

Ponadto do oceny dokładności zastosowano metrykę mAP (ang. *Mean Average Precision*), która mówi o dokładności modelu. Aby zapewnić kompleksową ocenę wydajności modelu w zestawieniu uwzględniona została również wartość mAP50, która oznacza średnią precyzję dla progu IoU wynoszącym 0.50 oraz wartość mAP50-95, która daje lepszy obraz na różnych poziomach trudności wykrywania dzięki obliczaniu przy różnych progach IoU w zakresie od 0.50 do 0.95.

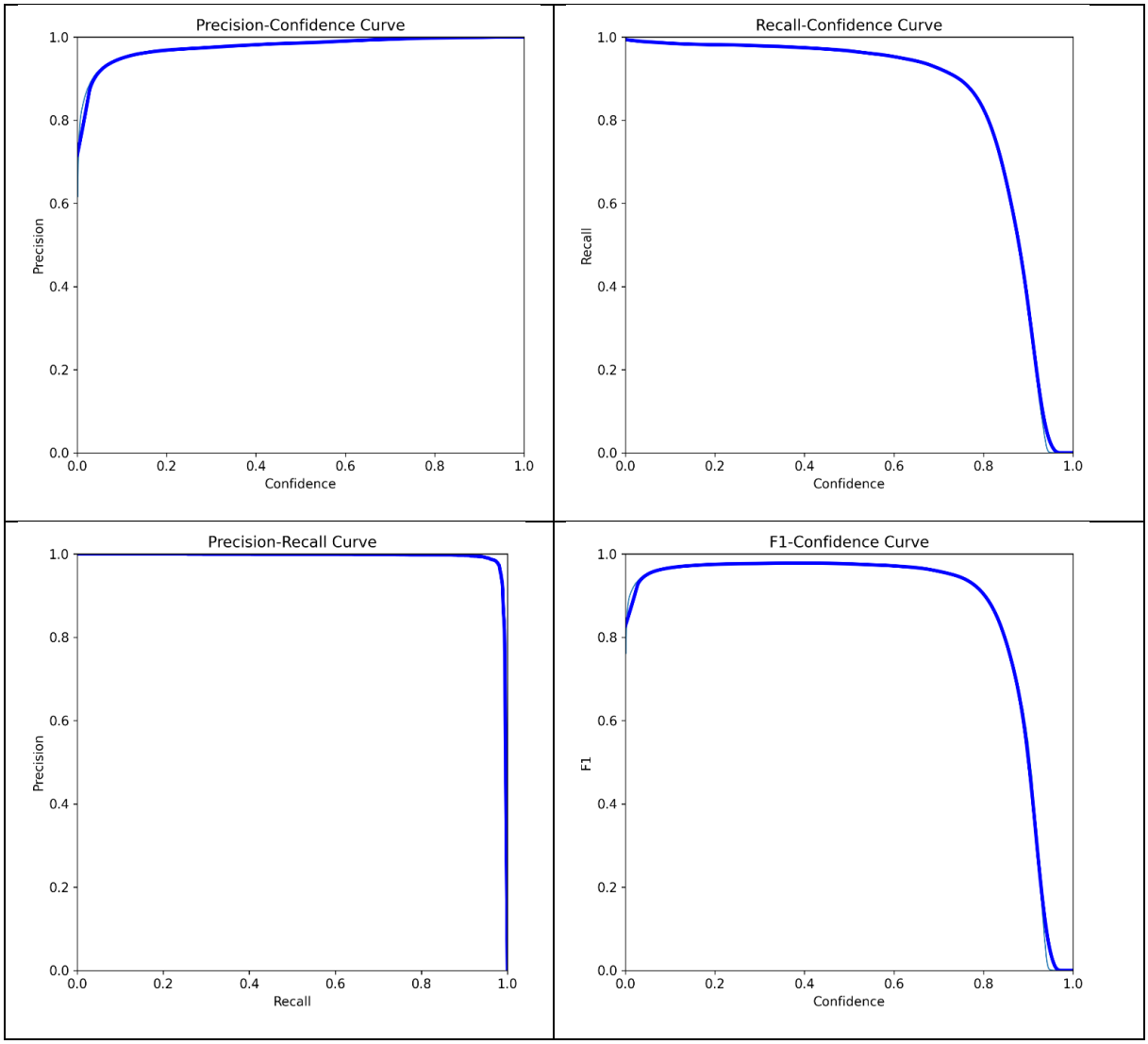
Cały proces uczenia modelu zakończył się po prawie 8 godzinach. Metryki dokładnościowe zostały zestawione poniżej (Tabela 17) z uwzględnieniem zbioru walidacyjnego i testowego.

Tabela 17 Zestawienie wyników uczenia modelu YOLOv8 – porównanie ze stanem prawdziwym dla zestawu walidacyjnego i testowego.

	Wyniki na zbiorze walidacyjnym (8672 zdjęć)	Wyniki na zbiorze testowym (10000 zdjęć)
Dokładność (precision)	0.986	0.981
Czułość (recall)	0.979	0.975
mAP50	0.993	0.992
mAP50-95	0.824	0.816

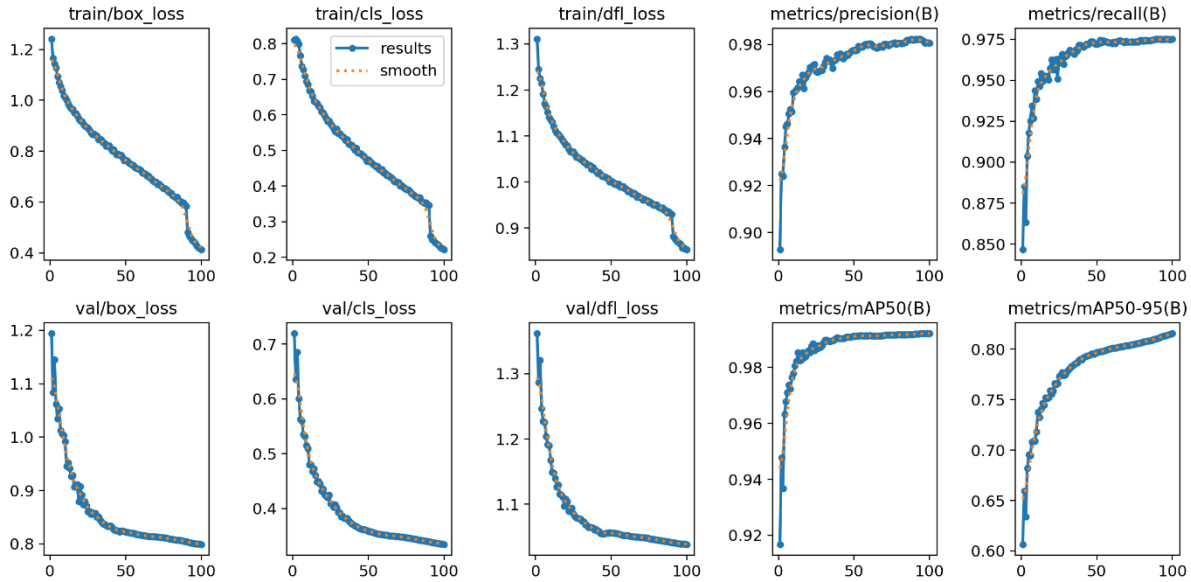
Analizując poniższe wykresy (Tabela 18), które w wizualny sposób ukazują dokładność modelu można powiedzieć, że precyzja pozostaje wysoka (blisko 1) dla większości poziomów ufności (*ang. confidence*), co wskazuje, że model dokonuje niewielu fałszywie pozytywnych predykcji i jest niezawodny w generowaniu predykcji o wysokim poziomie ufności. Krzywa *Recall-Confidence* ukazuje skuteczność modelu jako funkcję progu ufności. Zaczyna się blisko 1 i gwałtownie spada wraz ze wzrostem progu ufności. Model przewiduje większość obiektów przy niższych progach, jednak ten duży spadek jest zauważalny dopiero na poziomie ok. 0.8 wartości poziomu ufności. Krzywa *Precision-Recall* pokazuje równowagę między precyzją i czułością przy różnych progach. Wartość wskaźnika F1 jest średnią harmoniczną precyzji i czułości, zapewniając ocenę wydajności modelu z uwzględnieniem zarówno wyników fałszywie dodatnich, jak i fałszywie ujemnych.

Tabela 18 Wykresy ukazujące wgląd w metryki dokładnościowe wytrenowanego modelu YOLOv8 do detekcji latarni.



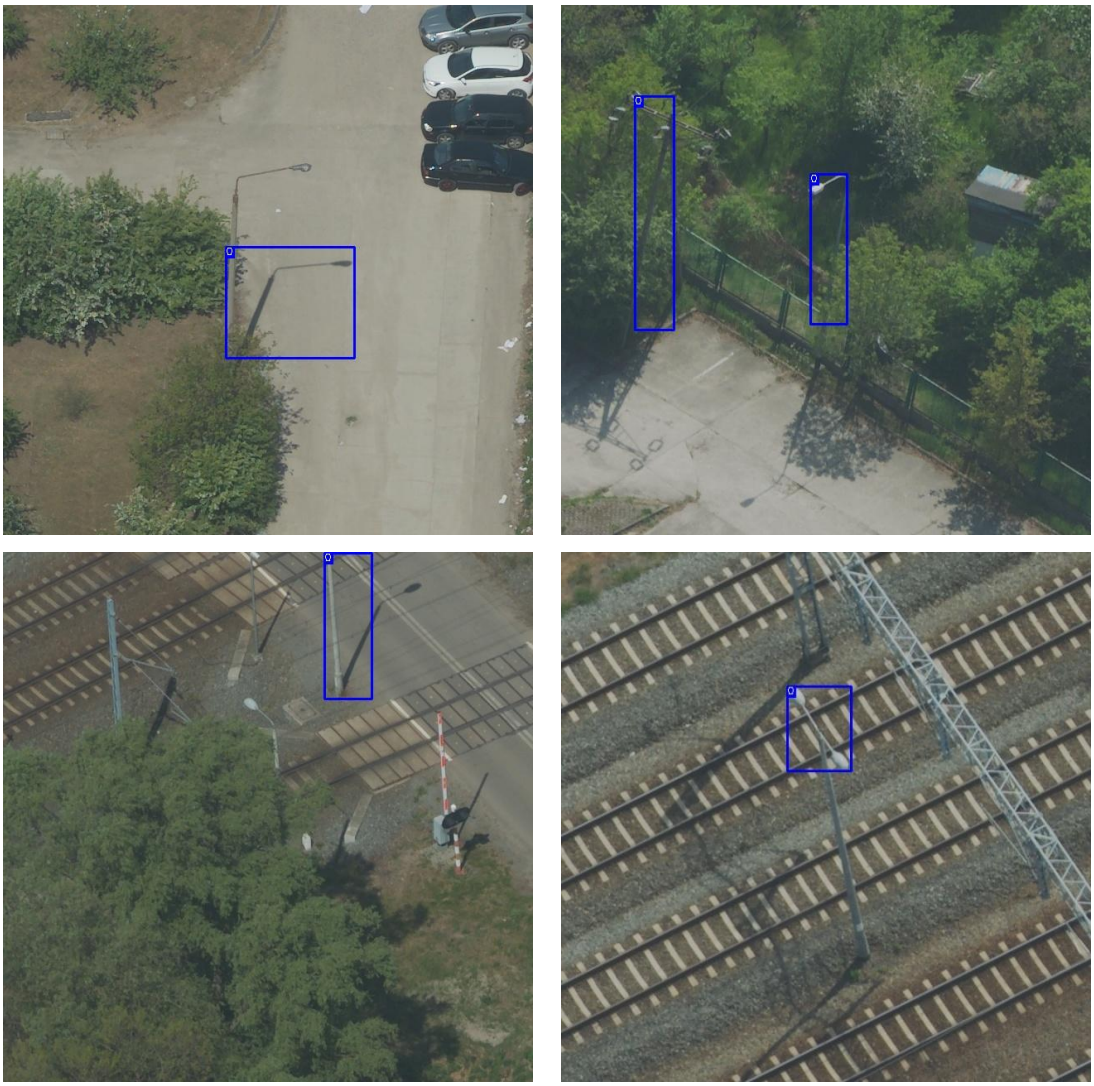
Ponadto na wykresach (Rysunek 16) pokazano dokładne metryki uzyskane podczas procesu uczenia. Rozdzielone zostały wartości dla zbioru uczącego oraz walidacyjnego. Interpretując krzywe można zaobserwować spadek wartości funkcji straty podczas treningu, co potwierdza fakt iż model z czasem lepiej radził sobie z detekcją latarni. Malejące wartości na wykresie „dfl_loss" wykazują, że podczas treningu model z lepszą precyzją określał położenie ramki ograniczającej. Ponadto można zauważyć, że wartości straty (*ang. loss*) dla zbioru walidacyjnego zbliżają się szybciej do zera niż odpowiednie wartości dla zbioru treningowego. Oznacza to, że model dobrze radził sobie z nieznanymi danymi i sugeruje, że szybko uczył się ważnych cech. Poprawa wskaźników wydajności oraz malejące wartości funkcji straty wraz z postępem treningu sugerują, że model uczył się skutecznie bez nadmiernego dopasowania. Podsumowując wyniki analizy dokładności treningu można powiedzieć, że model wykazuje

wysoką skuteczność w detekcji latarni z wysoką precyzją i czułością, zapewniając wiarygodne i dokładne wyniki.



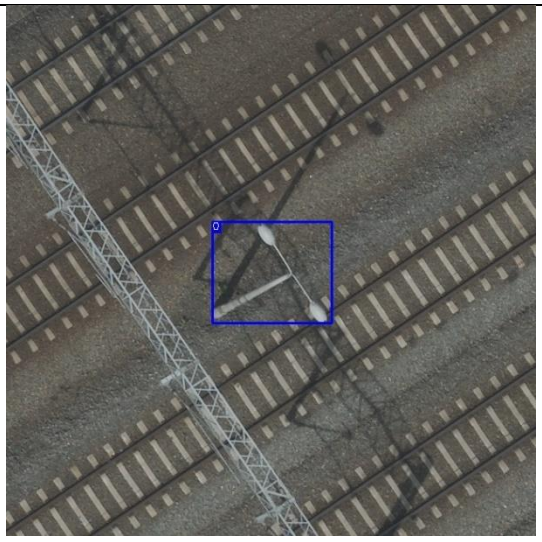
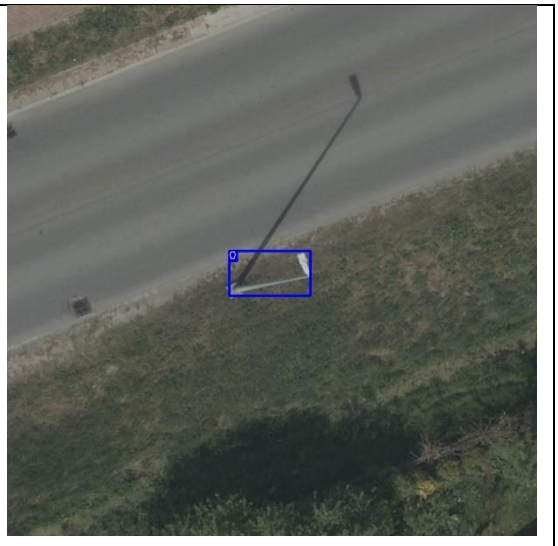
Rysunek 16 Wykresy metryk trenowania modelu YOLOv8 umożliwiające wgląd w wydajność modelu podczas całego procesu szkolenia.

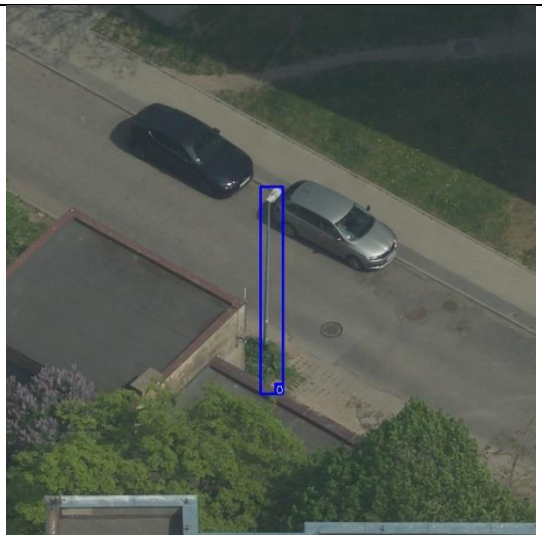
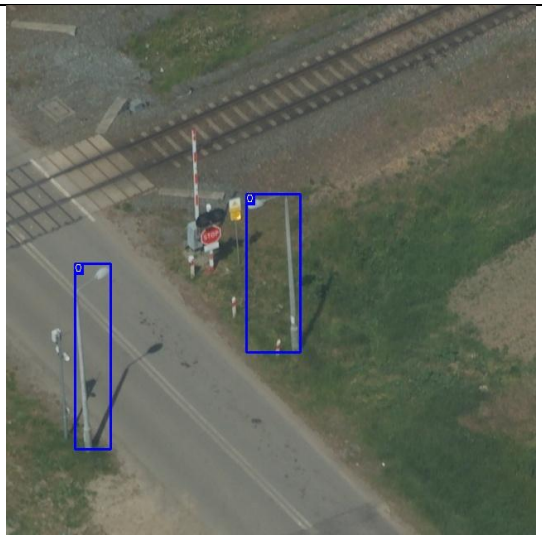
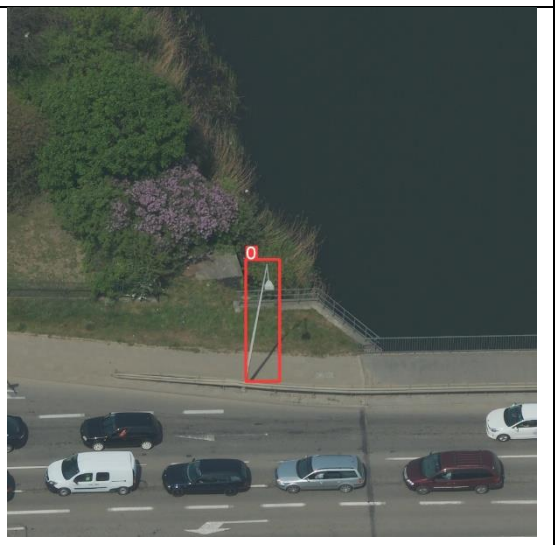
Oprócz analizy statystyk dokonano wizualnej oceny wyników. Do najczęstszych błędów, które występowały jeśli chodzi o detekcję latarni, było wykrywanie cieni latarni zamiast samej latarni, wykrywanie zarówno cienia, jak i latarni jako dwóch różnych obiektów, czy też detekcja innych obiektów, które nie były latarniami np. słupów energetycznych, które są trochę podobne jeśli chodzi o charakterystykę. Czasami model nie radził sobie z detekcją latarni częściowo przysłoniętych przez inne obiekty. Zdarzało się również tak, że w wyniku zaznaczono ramkę tylko dla górnej części obiektu, a nie dla całej latarni. Przykłady błędów pokazano poniżej (Rysunek 17).

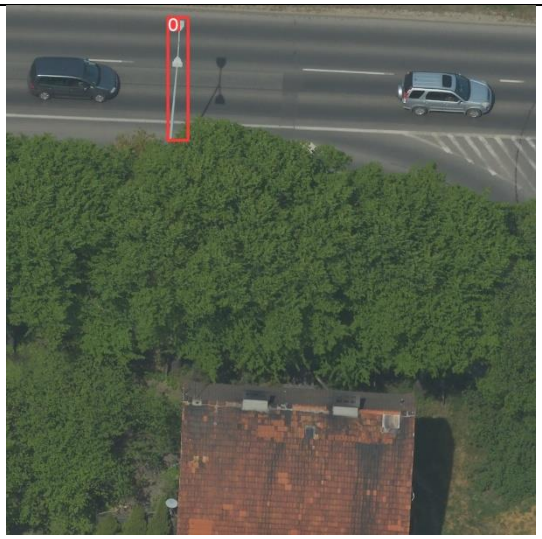
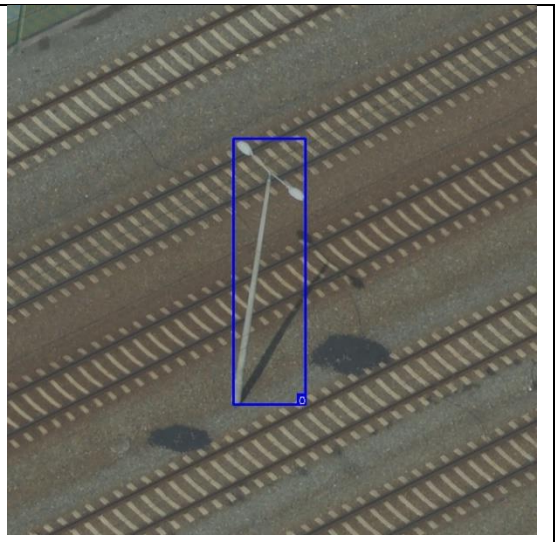


Rysunek 17 Przykłady błędów modelu YOLOv8 na zbiorze testowym.

Poniżej (Rysunek 18) pokazano kilka przykładowych poprawnych wyników detekcji latarni ze zbioru testowego dla czterech kierunków ukośnych oraz kierunku pionowego,

Nadir		

Left		
Right		
Forward		

Backward		

Rysunek 18 Przykładowe wyniki detekcji latarni wytrenowanym modelem YOLOv8 dla czterech kierunków ukośnych i kierunku nadir.

8. Ocena wyników modelu Segment Anything Model (SAM) dla latarni

Jak pokazały powyższe wyniki detekcja latarni na zdjęciach z wykorzystaniem modeli uczenia maszynowego wykazuje wysokie dokładności. Jedyną wadą tego podejścia jest to, że w rezultacie nie otrzymujemy dokładnego kształtu obiektu, a jedynie ramkę ograniczającą ten obiekt. Innym podejściem do wykrywania obiektów na zdjęciach jest ich segmentacja na obrazie, w wyniku której można uzyskać dokładny kształt obiektu. Oczywiście w zależności od zastosowania i oczekiwanego wyniku należy dobrać metodę.

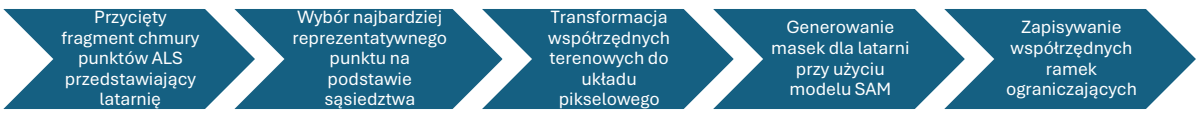
Model SAM (*Segment Anything Model*) został wybrany jako rozwiązanie w celu sprawdzenia, czy taki gotowy model segmentacji może obsługiwać wykrywanie latarni bez potrzeby szkolenia. Spośród dostępnych wag dla trzech różnych skal modelu Vision Transformer wybrano model ViT-B SAM. Model został przetestowany na laptopie wyposażonym w kartę graficzną NVIDIA GeForce RTX 3070. Początkowo sprawdzono, jaki wynik zostanie osiągnięty podczas segmentacji całego obrazu lotniczego bez podpowiedzi. Wcześniej jednak obrazy zostały podzielone na mniejsze kafelki, dzięki czemu obraz mógł zostać przepuszczony przez model. Przykład wyników segmentacji całego kafelka pokazano na Rysunek 19. Można zauważyć, że nie wszystkie elementy (w tym latarnie uliczne) zostały posegmentowane na obrazach. W szczególności duże problemy można było zauważyć dla wysokich, smukłych, liniowych obiektów.

Rysunek 19 Wyniki segmentacji całych kafelków obrazu za pomocą modelu SAM, gdzie zaznaczono brak wyodrębnionych latarni.

Ponieważ wyniki nie były satysfakcjonujące, nawet przy znacznie mniejszych kafelkach, zdecydowano się na inne podejście. Dzięki temu, że możliwe jest przekazanie modelowi podpowiedzi w postaci punktu, ramki ograniczającej lub tekstu, zdecydowano się wskazać punkt lokalizacji latarni na zdjęciu. Jest to możliwe poprzez kliknięcie na zdjęciu i wskazanie piksela obiektu, który ma zostać wyodrębniony z obrazu, albo poprzez podanie współrzędnych piksela. Aby pominąć ręczną pracę związaną z koniecznością klikania każdego zdjęcia, zastosowano wcześniejsze wyniki. Dla każdego obiektu (latarni) pobrano fragment chmury punktów ALS uzyskany we wcześniejszym kroku. Pojedynczy punkt chmury reprezentujący latarnię był wybierany i traktowany jako punkt wskazania dla latarni. Biorąc pod uwagę aspekty związane z kompletnością chmur punktów należało zastanowić się nad metodyką wyboru punktu, który następnie zostanie zrzutowany na zdjęcie, a współrzędne pikselowe będą stanowiły podpowiedź dla modelu.

8.1. Metodologia

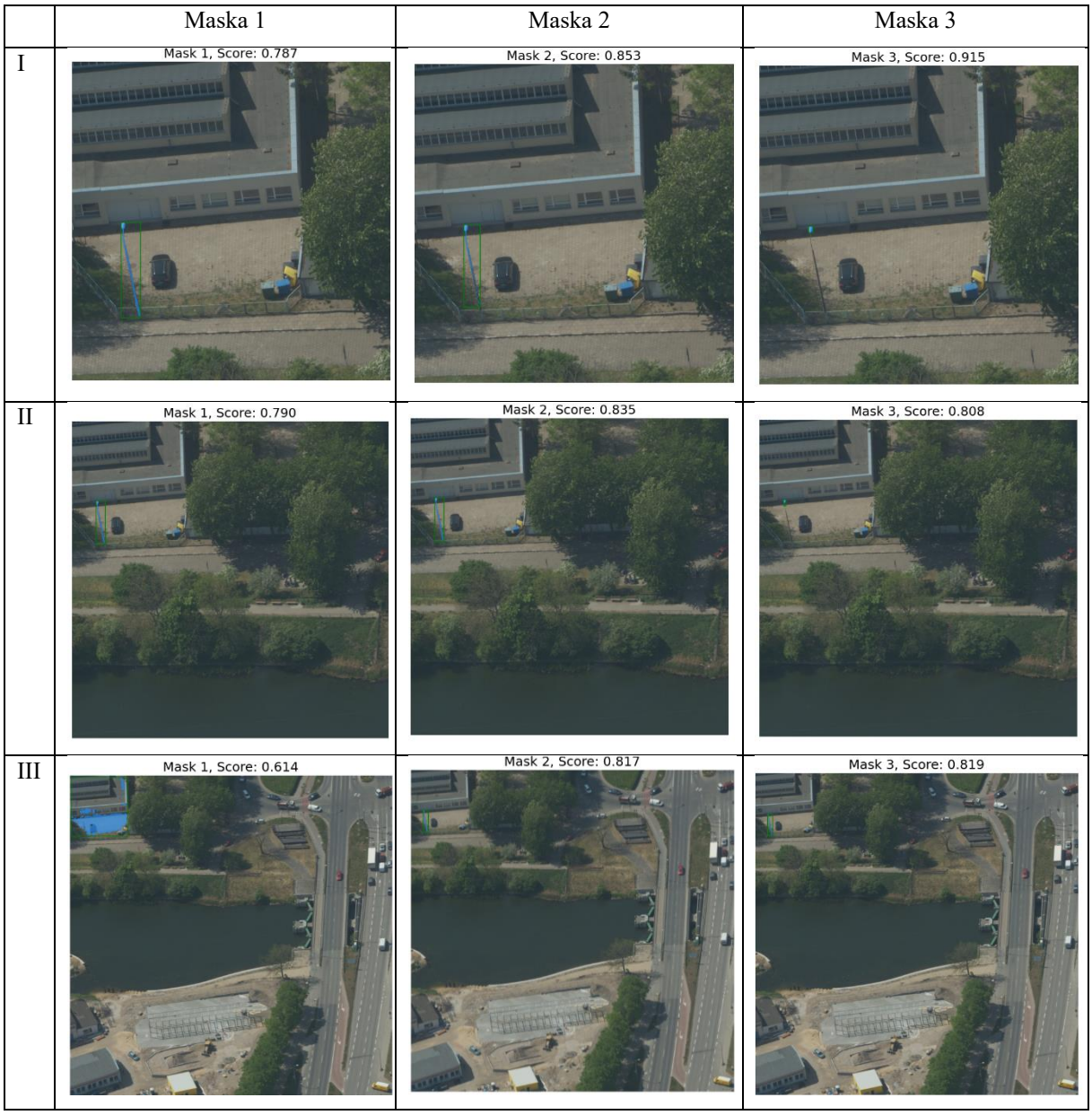
Zaproponowane podejście do wyboru punktów zakłada, że punkty znajdujące się powyżej połowy wysokości latarni ulicznej są najpierw odfiltrowywane. Umożliwia to wykluczenie wyboru punktu chmury w pobliżu gruntu. Dzięki temu wykluczona jest sytuacja, kiedy punkty należące do klasy niskiej roślinności lub gruntu mogły zostać wcześniej ewentualnie błędnie sklasyfikowane jako podstawa latarni. Następnie dla pozostałych punktów reprezentujących latarnię znajdujących się powyżej połowy wysokości policzono liczbę sąsiadów. Punkt o największej liczbie sąsiadów był punktem reprezentatywnym. W kolejnym kroku współrzędne wybranego punktu zostały przeniesione do układu pikselowego zdjęcia. Proponowana metodologia wyboru punktu chmury reprezentującego obiekt i wykorzystania go jako podpowiedzi dla modelu SAM została przedstawiona poniżej (Rysunek 20).



Rysunek 20 Kolejne etapy segmentacji latarni ulicznych z wykorzystaniem chmur punktów jako podpowiedzi do modelu SAM.

Po przekształceniu współrzędnych latarni do układu pikselowego, model wygenerował dla nich maski na tej podstawie. Jednak ze względu na ograniczenia obliczeniowe niemożliwe było dostarczenie do modelu obrazu w pełnej rozdzielczości. Aby zmniejszyć zbiór danych

i przyspieszyć proces, wokół piksela latarni ulicznej wycięto kafelek o rozmiarze 800x800 pikseli. Wybór rozmiaru kafelka wiązał się również z ograniczeniami zasobów obliczeniowych. Sprawdzono również wpływ rozdzielczości, czy przepróbkowanie obrazu do piksela 5 cm lub 10 cm poprawiłoby wyniki. Dla kilku przykładowych latarni zweryfikowano wpływ rozdzielczości na wynik, jak widać poniżej (Rysunek 21). Zmniejszenie rozdzielczości niestety negatywnie wpłynęło na wynik, więc ostatecznie wszystkie kolejne testy zostały przeprowadzone dla kafelków o rozmiarze 800x800 z oryginalną rozdzielczością.



Rysunek 21 Wynik segmentacji latarni (trzy różne maski o różnym poziomie ufności jako wynik z modelu) dla trzech różnych rozdzielczości I) dla oryginalnych 2,5 cm, II) po resamplingu do 5 cm, III) po resamplingu do 10 cm.

Ponadto jak widać powyżej domyślnie SAM generuje trzy maski, każda z innym wskaźnikiem poziomu zaufania (*ang. confidence score*) - tj. szacowanym IoU. Umożliwia to wybranie

najlepszej pojedynczej maski z najwyższym wynikiem. Poniżej znajduje się przykład tej samej latarni ulicznej na dwóch kolejnych zdjęciach z szeregu (Rysunek 22). Przedstawione zostały wyniki dla trzech masek. Jak widać na pierwszym obrazie (w pierwszym rzędzie), obwiednie dla maski nr 1 i nr 2 są bardzo podobne. W przypadku maski nr 1 i nr 2 SAM uwzględnił również cień latarni ulicznej jako ten sam obiekt. W przypadku maski nr 3 z najwyższym wynikiem, maska ogranicza tylko górną część latarni (oprawę). Podobny wynik dla maski nr 3 uzyskano dla drugiego obrazu (drugi wiersz w tabeli). Jednak w przypadku maski nr 2, chociaż wartość wyniku zaufania nie jest najwyższa, cały obiekt został poprawnie wykryty, bez dodatkowych części obrazu.

Rysunek 22 Wyniki dla jednej latarni ulicznej z modelu SAM dla dwóch zdjęć z szeregu dla trzech masek o różnej wartości poziomu zaufania.

Ponieważ wyniki dla modelu SAM zostały zapisane dla kafelków o rozmiarze 800x800, konieczne było ponowne obliczenie współrzędnych ramek ograniczających w odniesieniu do całego rozmiaru obrazu, aby móc porównać wyniki dla całego obrazu w oryginalnej rozdzielczości. Poniższy Rysunek 23 przedstawia przykładowy wynik porównania trzech ramek ograniczających wygenerowanych z masek o różnych wynikach zaufania przeniesionych na ukośne zdjęcie lotnicze w pełnej rozdzielczości.

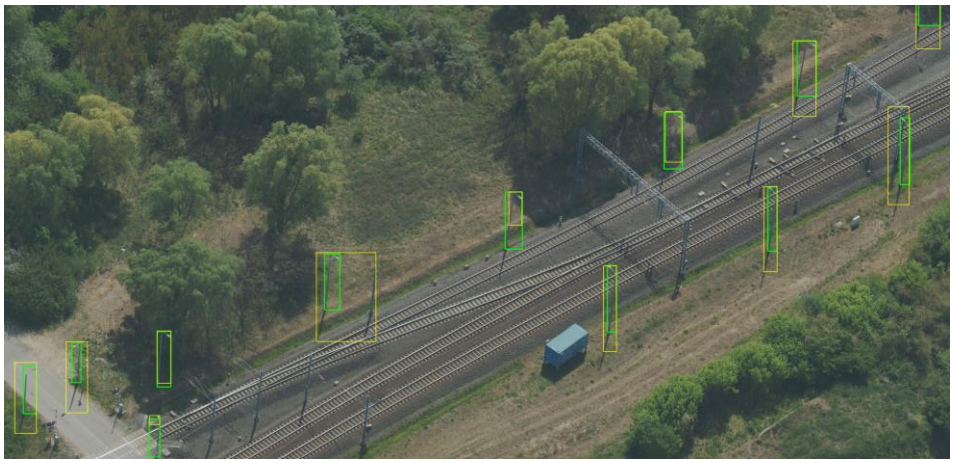
Rysunek 23 Wyniki dla modelu SAM dla trzech różnych masek z różnymi wynikami zaufania. Kolor żółty odpowiada masce o najmniejszej wartości współczynnika „confidence score”, kolor czerwony średniej, a kolor zielony najwyższej wartości.

8.2. Porównanie modelu SAM z referencyjnym zestawem danych testowych

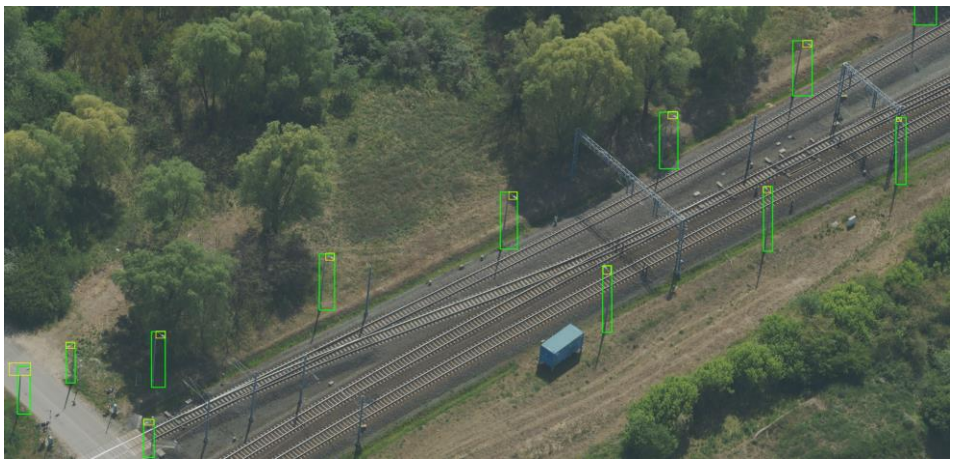
Wstępna ocena wizualna wykazała, że model SAM, mimo że nie został przeszkolony na dodatkowym zbiorze danych, jest w stanie dokonać segmentacji latarni na obrazie po wskazaniu piksela, który został obliczony na podstawie transformacji wybranego punktu chmury ALS do układu współrzędnych pikselowych zdjęcia. Niestety, widoczne były problemy związane głównie z cieniami rzucanymi przez obiekt. Piksele te zostały uwzględnione w segmencie obiektu. W związku z tym, mimo że latarnia uliczna została wykryta w danej lokalizacji, obwiednia wygenerowana na podstawie masek często różniła się od danych referencyjnych. Innym przykładem niedokładnego wyniku segmentacji była sytuacja, gdy obiekt „kładł się” na zdjęciu na inny podobny do latarni. Na przykład gdy na obrazie znajdowało się ogrodzenie lub inny obiekt o podobnej charakterystyce. W takim wypadku zdarzało się, że oprócz latarni do segmentu zaliczane były również piksele innych obiektów. Poniżej przedstawione zostały przykładowe wyniki dla każdego z kierunków dla trzech masek o zróżnicowanym poziomie zaufania. Dla każdego z kierunków: wstecz (Rysunek 24), w przód

(Rysunek 25), pionowo (Rysunek 26), w lewo (Rysunek 27), w prawo (Rysunek 28) porównano ramki ograniczające latarnie, które powstały w wyniku segmentacji przy użyciu SAM z danymi referencyjnymi.

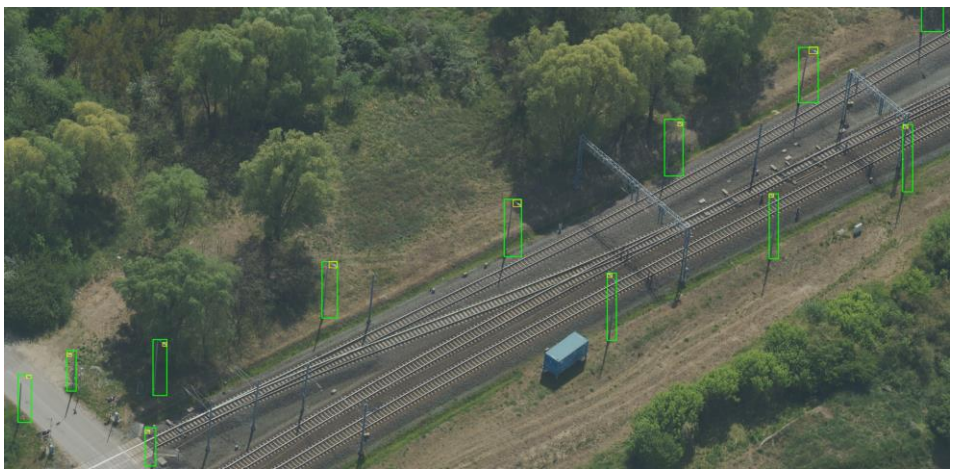
1



2

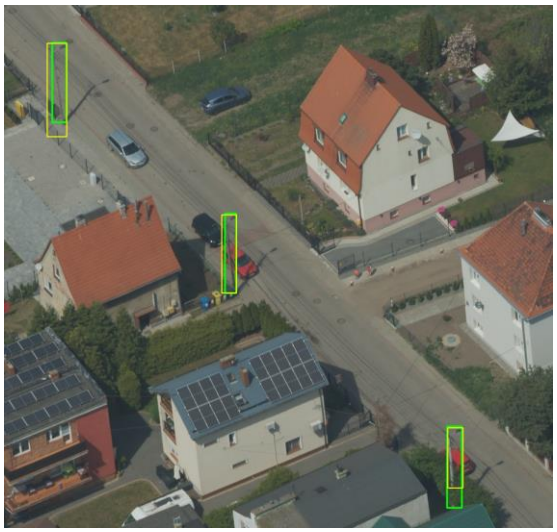


3

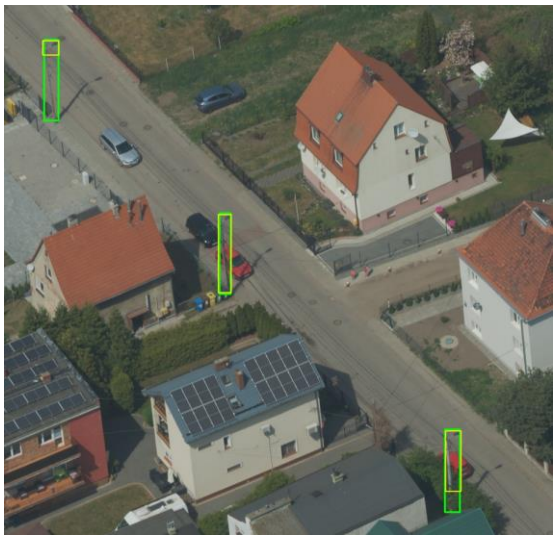


Rysunek 24 Wyniki dla modelu SAM (kolor żółty) dla trzech różnych masek (uszeregowane kolejno od góry od najmniejszej wartości zaufania do najwyższej). Zielone obramowanie to referencja, a żółte to wynik predykcji (kierunek wstecz - backward).

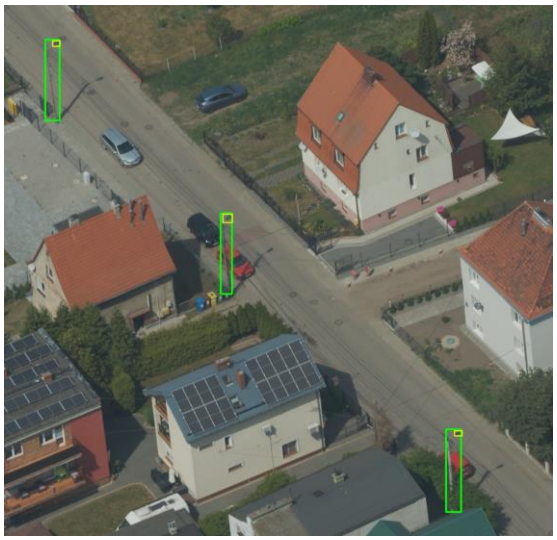
1



2



3



Rysunek 25 Wyniki dla modelu SAM (kolor żółty) dla trzech różnych masek (uszeregowane kolejno od góry od najmniejszej wartości zaufania do najwyższej). Zielone obramowanie to referencja, a żółte to wynik predykcji (kierunek w przód - forward).

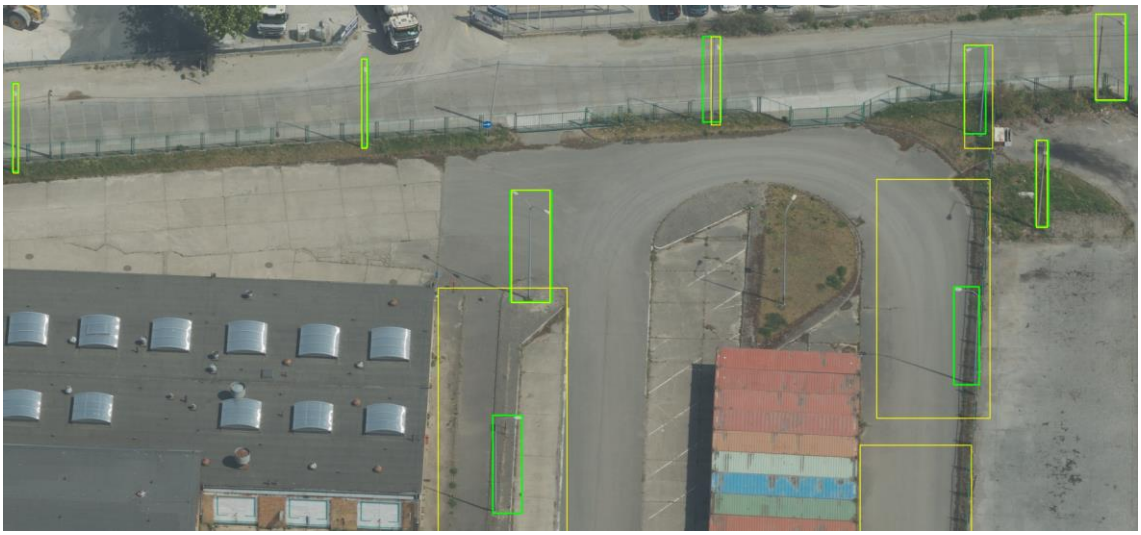
1

2

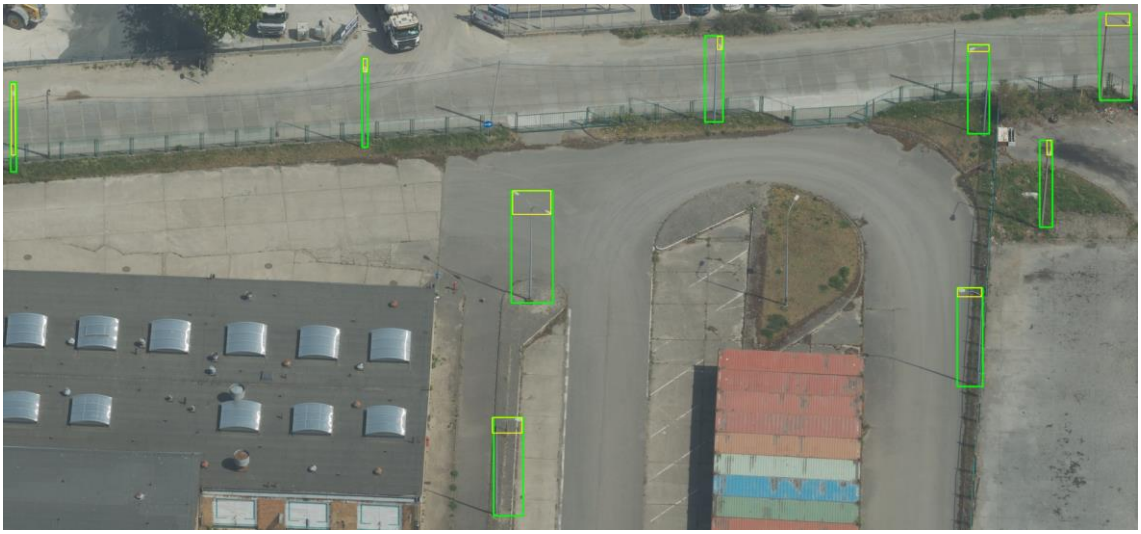
3

Rysunek 26 Wyniki dla modelu SAM (kolor żółty) dla trzech różnych masek (uszeregowane kolejno od góry od najmniejszej wartości zaufania do najwyższej). Zielone obramowanie to referencja, a żółte to wynik predykcji (kierunek pionowy - nadir).

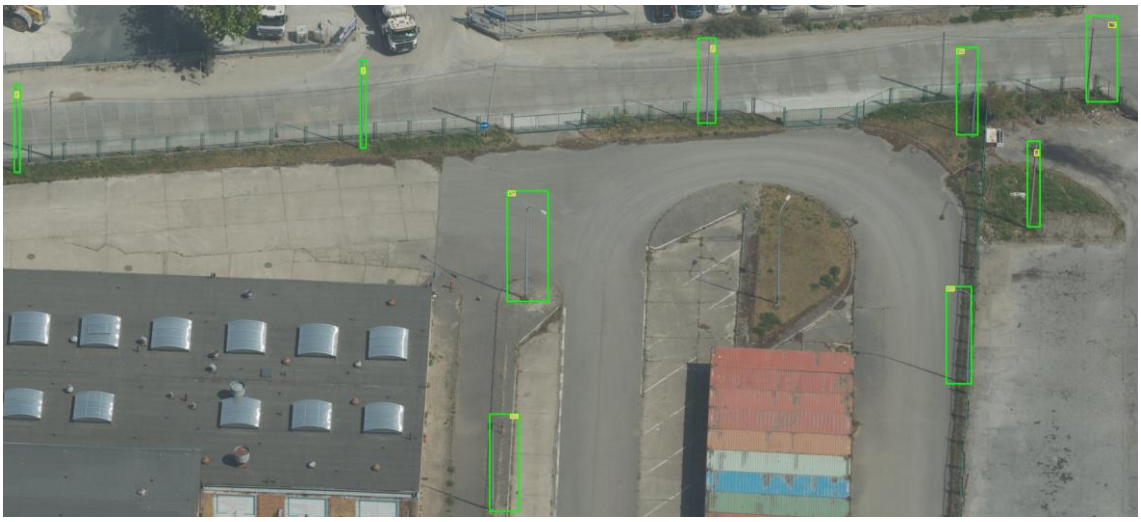
1



2

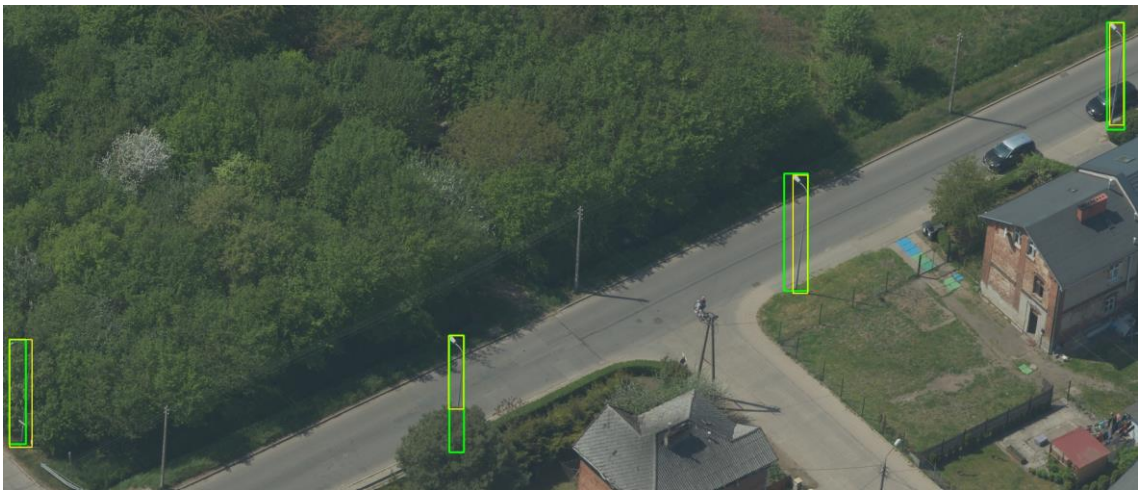


3

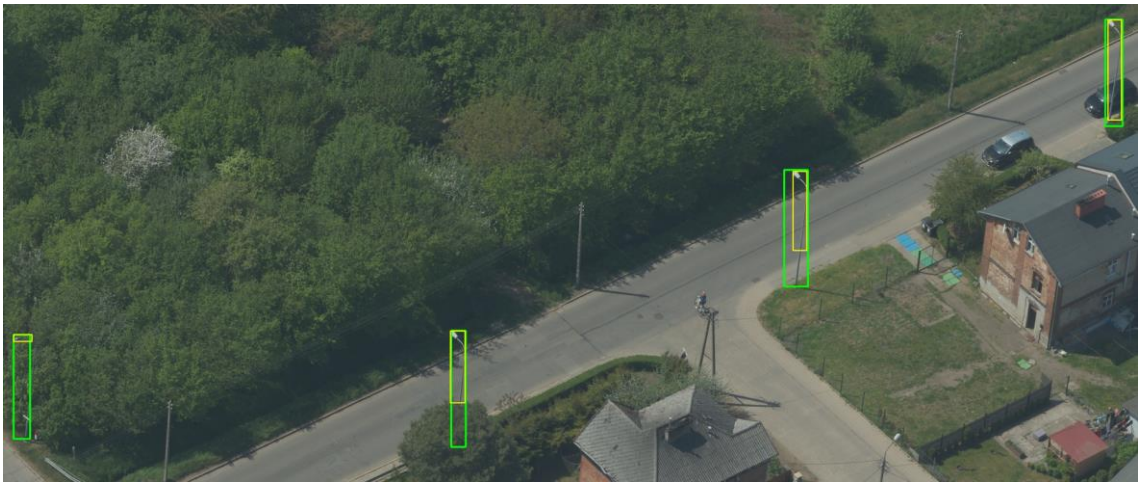


Rysunek 27 Wyniki dla modelu SAM (kolor żółty) dla trzech różnych masek (uszeregowane kolejno od góry od najmniejszej wartości zaufania do najwyższej). Zielone obramowanie to referencja, a żółte to wynik predykcji (kierunek w lewo - left).

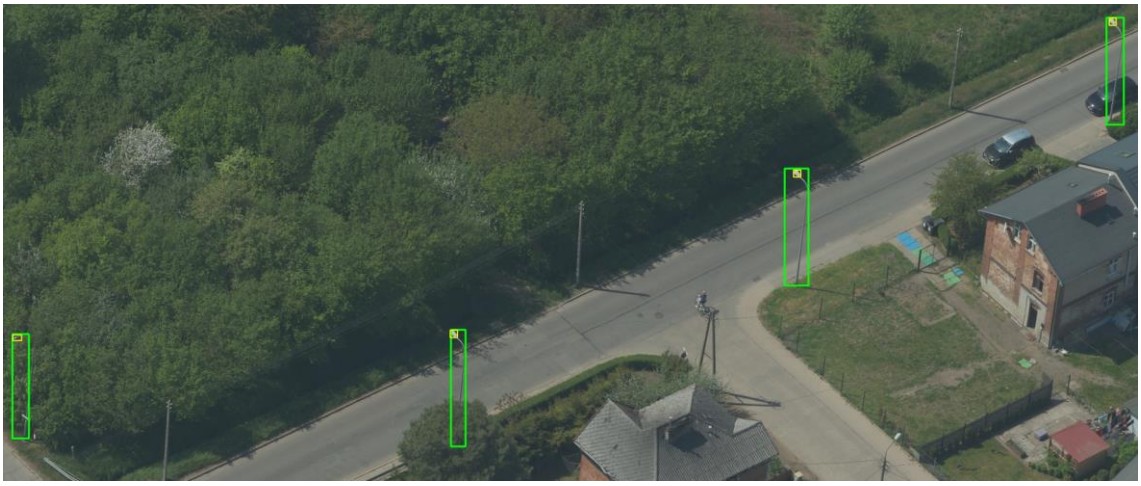
1



2



3

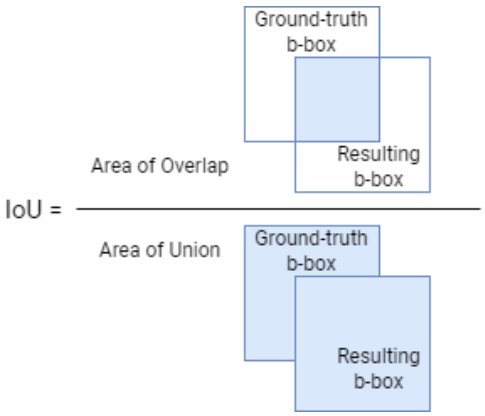


Rysunek 28 Wyniki dla modelu SAM (kolor żółty) dla trzech różnych masek (uszeregowane kolejno od góry od najmniejszej wartości zaufania do najwyższej). Zielone obramowanie to referencja, a żółte to wynik predykcji (kierunek w prawo - right).

Dane referencyjne, jak wspomniano, to ramki ograniczające obliczone na podstawie konwersji chmury punktów ALS na ukośne zdjęcia lotnicze. Jak widać na powyższych rysunkach, po analizie wizualnej można stwierdzić, że maski nr 3 z najwyższym wynikiem zaufania w większości zawsze oznaczały tylko oprawę latarni ulicznej. Wynika to prawdopodobnie z faktu, że punkty wybrane automatycznie do wskazania obiektu na zdjęciu znajdowały się

najczęściej w górnej części obiektu, gdyż gęstość chmury punktów w tej części latarni była wyższa w porównaniu punktów na słupie czy wysięgniku.

Chociaż SAM nie zawsze w pełni poprawnie wysegmentował obiekt, można zauważyć, że w przypadku masek nr 1 i nr 2 wyniki były lepsze lub gorsze, ale pokrywały się z danymi referencyjnymi. Dzięki temu, że w rezultacie dla każdej latarni na zdjęciu uzyskano trzy różne maski, możliwe było wybranie spośród nich tej, która była najbardziej podobna do referencyjnej ramki ograniczającej. Jako miarę dokładności wykorzystano wartość wskaźnika Intersection over Union (IoU). IoU służy do oceny dokładności detektorów obiektów na danym zbiorze danych. Jako dane referencyjne (*ang. gound-truth*) wykorzystano ramki ograniczające (*ang. bounding-boxy*) z referencyjnego zbioru danych opartego na chmurach punktów ALS. Predykcje natomiast to dane wyjściowe z modelu, które były obwiedniami segmentów latarni. Metryka IoU odnosi się do oceny dokładności nakładania się danych wyjściowych z modelu w porównaniu z danymi referencyjnymi, gdzie licznik jest obszarem wspólnym pomiędzy ramką przewidzianą przez model a ramką rzeczywistą (*ang. ground-truth*). Mianownik natomiast to suma obszarów tych dwóch ramek ogarniających (Rysunek 29).



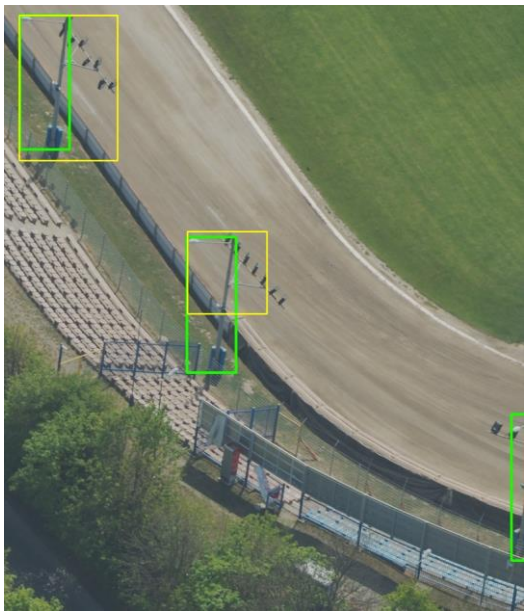
Rysunek 29 Intersection over Union.

Wszystkie wynikowe ramki ograniczające dla 1959 obrazów zostały porównane z danymi referencyjnymi. Ze względu na fakt, że dla każdego obiektu SAM wygenerował trzy maski, do oceny dokładności wybrano tę, dla której wartość IoU była najwyższa. Średnia wartość IoU, która określa dokładność wyniku z modelu SAM względem referencji została określona dla każdego kierunku (Tabela 19).

Tabela 19 Porównanie dokładności dla pięciu kierunków - porównanie wyników z modelu SAM z referencjami.

Kierunek	Liczba wykryć	Liczba latarni niewykrytych przez model SAM	IoU [%]
Nadir	6429	184	61,50
Forward	9789	404	60,58
Backward	10249	389	49,34
Left	7806	338	53,84
Right	9357	307	53,21

Jak widać w tabeli powyżej (Tabela 19), najwyższa dokładność została osiągnięta dla kierunku pionowego, dla którego średnia wartość IoU wyniosła 61,50%, a najgorsza dla kierunku *backward* - 49,34%. Tabela pokazuje również liczbę obiektów, których SAM w ogóle nie wykrył (tzn. nie została utworzona przez model żadna maska dla wskazanego piksela). W dalszej analizie wyników sprawdzono, które latarnie uliczne nie zostały wykryte. Zauważono, że największy problem występował na krawędziach obrazów, gdzie tylko część latarni była widoczna na obrazie. Pomimo, że takie obiekty na krawędzi występowały w danych referencyjnych, nawet jeśli nie cała latarnia była widoczna, to w przypadku predykcji, gdy punkt, który został wybrany jako wejście, podpowiedź do modelu i znajdował się poza obrazem zdjęcia po konwersji na współrzędne pikseli naturalnie SAM nie wykrywał niczego na obrazie. Poniżej przedstawiono kilka przykładów brakujących detekcji na krawędziach obrazu (Rysunek 30).



Rysunek 30 Przykłady pokazujące problemy z wykrywaniem latarni na krawędziach obrazów.

Potencjał wykorzystania gotowych modeli głębokiego uczenia do segmentacji i wykrywania obiektów na obrazach jest oczywisty. W szczególności w przypadkach takich jak wykrywanie latarni na ukośnych zdjęciach lotniczych, gdzie nie ma dostępnych szkoleniowych zbiorów danych. Czasochłonne ręczne tworzenie takich zbiorów danych można zastąpić wykorzystaniem modeli takich jak SAM, które zostały przeszkolone na znacznej ilości danych szkoleniowych przez duże firmy. Biorąc pod uwagę, że model osiągnął dokładność 50-60% przy domyślnych ustawieniach bez ponownego trenowania i uaktualniania wag, można to uznać za całkiem dobry wynik. Jednakże dla takich obiektów jak latarnie, które charakteryzują się tym, że są smukłym obiektem, lepszym wyborem jest detekcja niż segmentacja, co potwierdzają powyższe wyniki.

9. Segmentacja semantyczna ukośnych zdjęć lotniczych z wykorzystaniem modelu SAM dla detekcji okien

Uzasadnienie wyboru okien jako drugiego obiektu badawczego do eksperymentów zostało szczegółowo opisane w rozdziale 4. Jednakże warto podkreślić fakt, iż wraz z rosnącym zapotrzebowaniem na wysoki poziom szczegółowości modeli budynków, ważna jest nie tylko geometria, ale także semantyka elementów elewacji. Okna stanowią zaś jeden z najważniejszych elementów budynku. Dlatego też dokładna ich ekstrakcja na podstawie naturalnych scen naziemnych czy ukośnych zdjęć lotniczych budzi zainteresowanie wśród badaczy w różnych zastosowaniach takich jak inspekcje termiczne (Kim i in., 2016) czy ocena ryzyka powodziowego (Amirebrahimi i in., 2016).

Do wykrywania okien wykorzystuje się różne typy danych, nie tylko obrazy, ale również chmury punktów. Jednakże podejściem, które zostało wybrane do rozpoznawania okien w pracy była segmentacja obrazów. W tym konkretnym przypadku zastosowania segmentacja wydała się korzystnym rozwiązaniem, gdyż umożliwiła bardziej precyzyjne określenie kształtu i granic obiektu niż w przypadku detekcji. Ponadto wykorzystanie zdjęć ukośnych umożliwiło widok z takiej perspektywy, gdzie fasada z oknami jest dobrze odwzorowana.

W badaniach zdecydowano się na wykrywanie okien na lotniczych zdjęciach ukośnych przez opartą na głębokim uczeniu semantyczną metodę segmentacji. Wybrano do tego zadania ponownie model do segmentacji obrazów – Segment Anything Model (SAM). Ta część eksperymentów miała na celu sprawdzenie potencjału wykorzystania modelu segmentacji bez trenowania od początku do wykrywania okien na lotniczych zdjęciach ukośnych.

Podejściem, które zostało wykorzystane w badaniach było zastosowanie wytrenowanych wag SAM, które autorzy udostępnili wraz z kodem źródłowym. Model SAM może być wyposażony w 3 różne enkodery. Na potrzeby badań przetestowano w pierwszej iteracji dwa modele ViT-B oraz ViT-H, które różnią się liczbą parametrów, natomiast dla drugiego obszaru badawczego wykorzystano tylko jeden model, który okazał się lepszy. Badania były wykonywane na sprzęcie wyposażonym w NVIDIA GeForce RTX 3070.

Proces segmentacji obejmował zastosowanie modelu SAM do każdego obrazu w celu wygenerowania masek. Do automatycznego generowania masek wykorzystano funkcję *SamAutomaticMaskGenerator.* W wyniku działania narzędzia generowana była lista słowników opisujących poszczególne segmentacje.

Pomimo tego, że model posiada funkcje filtracji masek, usuwania duplikatów i zachowywania tylko najlepszych dzięki algorytmowi tłumiącemu NMS (*ang. non-maxiumum suppression*), w efekcie uzyskiwano maski dla całości obrazu (Rysunek 31), nie tylko dla okien. Ponadto wynikowe maski nie zawierały informacji o przypisanej klasie. Konieczne więc było odrzucenie wszystkich innych i pozostawienie tylko tych przedstawiających okna. Aby łatwiej było pracować z wynikami przetransformowano maski do postaci poligonów w formacie *.shp.

Rysunek 31 Wynik segmentacji obrazu z wykorzystaniem modelu SAM, gdzie każdy kolor oznacza oddzielną maskę.

Czyszczenie zestawu masek obejmowało kilka kroków: usuwanie zbyt małych i zbyt dużych masek, a także łącznie masek, które się nakładały. Wykorzystano do tego m.in. funkcję *dissolve* z zestawu narzędzi ArcGIS Pro, która zagregowała pokrywające się maski. Predykcje zostały następnie porównane z zestawem masek referencyjnych w celu oceny wydajności modelu.

Jak wspomniano, SAM segmentuje każdy obiekt na obrazie, generując odrębne maski dla każdej instancji bez zapewnienia informacji o typie obiektu. Naprzeciw temu ograniczeniu ESRI opracowała model Text SAM integrujący SAM z Grounding Dino. Grounding DINO to otwarto-źródłowy model języka wizyjnego, który radzi sobie z wykrywaniem obiektów na obrazach na podstawie ich opisu tekstowego. W ramach pakietu głębokiego uczenia oba modele są wywoływane sekwencyjnie. Dzięki takiemu podejściu podpowiedzi tekstowe dostarczone przez użytkownika wspierają wyodrębnienie obiektów zainteresowania.

Potencjał widoczny w rozwiązaniu był motywacją do podjęcia próby wykorzystania modelu do zadania segmentacji okien. Jednakże kilkukrotne próby z wykorzystaniem różnych parametrów nie przyniosły poprawnych wyników. Prawdopodobnie Text SAM nie jest jeszcze

dostosowany pod dane o różnej charakterystyce, gdyż jest stosukowo nowym modelem udostępnionym na początku lutego 2024 i powinien zostać jeszcze dopracowany.

9.1. Dane i obszar badawczy

Jak wspomniano, pomimo, że okna wydają się być dość prostym obiektem do wykrywania, to istnieje kilka wyzwań, które są widoczne w tym zadaniu. Różnorodność stylów budynków skutkuje złożonością geometrii okien. Co więcej dekoracje elewacji, które wyglądają podobnie do okien mogą powodować błędy. Okiennice, zacienienie, odbicia szła również mogą wpływać na wyniki. Z tego też względu wykorzystano dwa różne obszary badawcze, które odznaczają się różną charakterystyką jeśli chodzi o obiekt zainteresowania.

Bordeaux

Pierwszym zestawem danych wykorzystanym na potrzeby badań były wysokorozdzielcze zdjęcia ukośne pozyskane sensorem Leica CityMapper o rozdzielczości 5cm. Obszar, dla którego pozyskano dane to centrum miasta Bordeaux we Francji. Ze względu na inne podejście badawcze w metodyce wykrywania okien w porównaniu z latarniami wykorzystujące wytrenowany model do segmentacji bez dodatkowego treningu, nie było konieczności podziału zestawu na treningowy, testowy i walidacyjny. Baza danych została utworzona manualnie z wykorzystaniem narzędzia do tworzenia adnotacji na obrazach na potrzeby projektu wdrożeniowego realizowanego na zlecenie firmy OPEGIEKA Sp. z o.o. w Zakładzie Fotogrametrii, Telelekcji i Systemów Informacji Przestrzennej Politechniki Warszawskiej (2020). Jeśli chodzi o utworzony zbiór, definicja okna była ściśle określona, aby baza danych tworzyła spójny zbiór. Mimo różnorodności w charakterystyce (Rysunek 32), obiekty te nie były dzielone na różne rodzaje.

Rysunek 32 Przykłady budynków z różnymi oknami w mieście Bordeaux.

Dane posłużyły jako referencja do oceny wyników i analizy segmentacji modelem SAM. Zbiór testowy zawierał 12 816 okien, a przykłady z bazy danych widoczne są poniżej (Rysunek 33).

Rysunek 33 Przykładowe dane referencyjne.

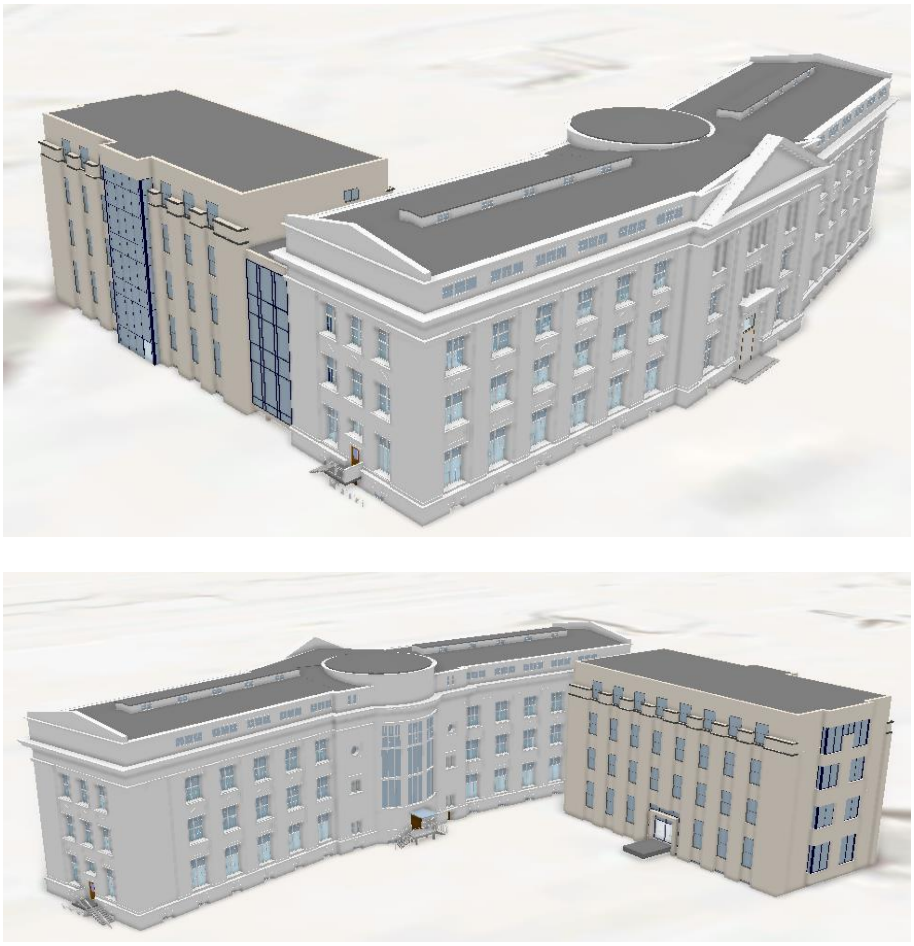
Pełne zdjęcia ukośne o rozdzielczości 10336 × 7788 pikseli wymagałyby zbyt dużych zasobów pamięci operacyjnej, żeby mogły być bezpośrednio przetwarzane przez model. Z tego powodu konieczne było ich podzielenie na mniejsze kafelki o rozmiarze 800x800 pikseli z pokryciem 50%. Ostatnim etapem procesu po wykryciu i wyodrębnieniu poprawnych masek okien konieczne było połączenie wyników dla każdego kafelka w jeden duży obraz o źródłowej rozdzielczości.

Warszawa

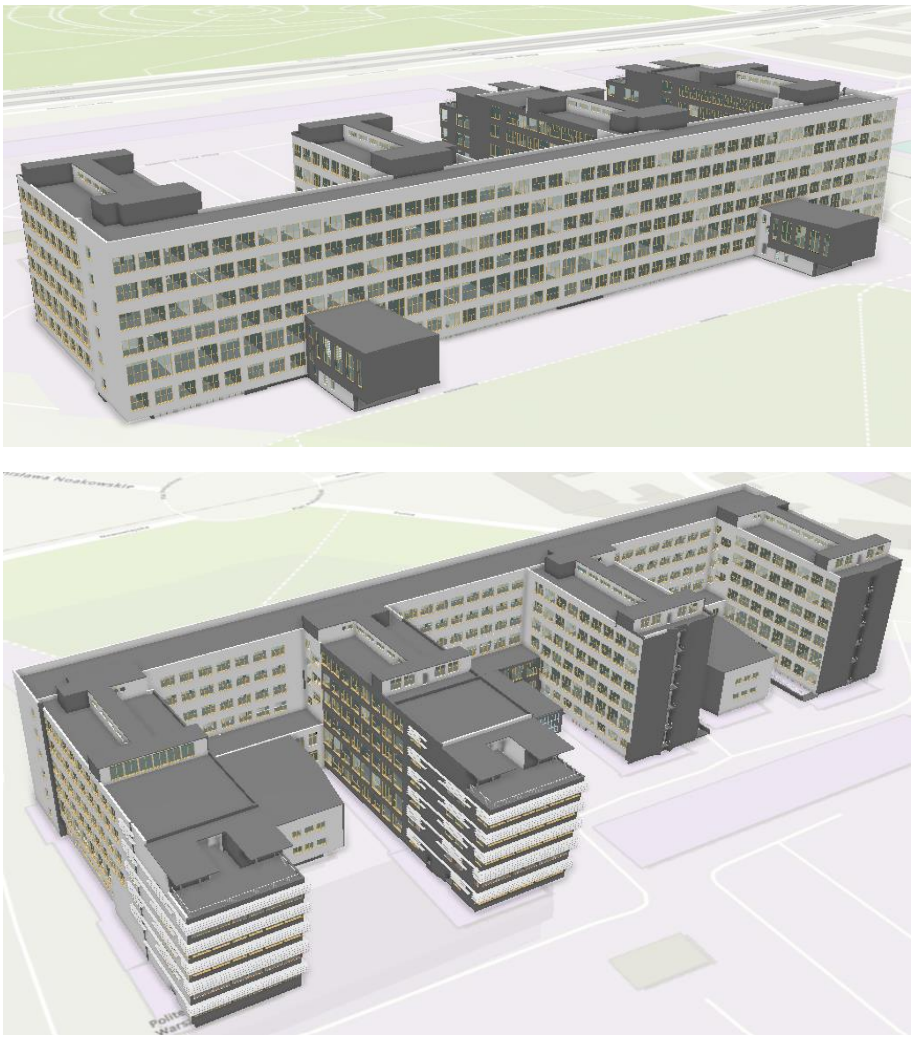
Na potrzeby eksperymentów wykorzystano również dane z tego samego sensora, ale dla innego obszaru – Warszawy. W przypadku danych dla Warszawy nie został utworzony zbiór referencyjny w manualny sposób tak, jak w przypadku Bordeaux ze względu na czasochłonność procesu. Jednakże na potrzeby projektu pt. „Politechnika Warszawska Ambasadorem Innowacji na Rzecz Dostępności" finansowanego z Programu Operacyjnego Wiedza Edukacja Rozwój 2014-2020 powstały modele BIM (*ang. Building Information Modeling*) budynków.

Ze względu na to, iż okna stanowiły jedną z rodzin całego modelu możliwe było wyodrębnienie obiektów zainteresowania. Oczywiście utworzenie spójnego zbioru testowego wymagało wykonania kilku kroków.

Po pierwsze spośród budynków Politechniki Warszawskiej wybrano dwa budynki, które różniły się jeśli chodzi o styl. Drugim kryterium wyboru było jak najmniejsza ilość drzew i innych obiektów, które przysłaniały budynek. Zdecydowano się na dwa obiekty: budynek Wydziału Transportu (Rysunek 34) oraz budynek Wydziału Elektroniki i Technik Informacyjnych (EITI) Politechniki Warszawskiej (Rysunek 35).

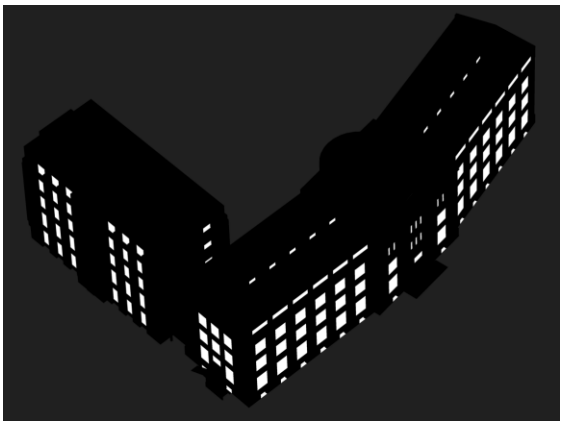
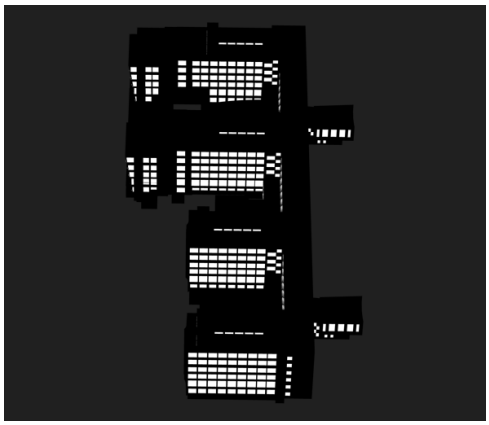


Rysunek 34 Model BIM budynku Wydziału Transportu Politechniki Warszawskiej.



Rysunek 35 Model BIM budynku Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej.

Aby możliwe było porównanie obrazu wynikowego z segmentacji z referencją konieczne było przetransformowanie modelu do postaci rastra. Z modelu BIM dla obu budynków wybrane zostały tylko dwie klasy, fasady oraz okna. Ze względu na to, że modele były w układzie lokalnym konieczne było nadanie im georeferencji. Następnie modele w odpowiednim formacie i układzie odniesienia zostały zaimportowane do projektu w programie Agisoft Metashape. Dzięki funkcji *Model.renderImage* możliwe było wygenerowanie widoków modelu na obraz 2D dla pozycji kamery, dzięki znanym elementom orientacji wewnętrznej i zewnętrznej zdjęć. W efekcie uzyskano obrazy, które posłużyły do porównania z wynikami SAM. Przykładowe wyrenderowane widoki modelu zostały przedstawione poniżej (Rysunek 36).

a)  b) 

Rysunek 36 Przykładowe wyniki renderowania – przetworzenia widoku modelu 3D do obrazu 2D a) Wydział Transportu, b) Wydział EITI.

Dla obu budynków wybrane zostało po 30 zdjęć z czterech kierunków ukośnych. Cały zbiór składał się z 1202 okien, z czego większa część to były okna budynku EITI (Tabela 20).

Tabela 20 Podstawowe informacje o zestawie danych referencyjnych.

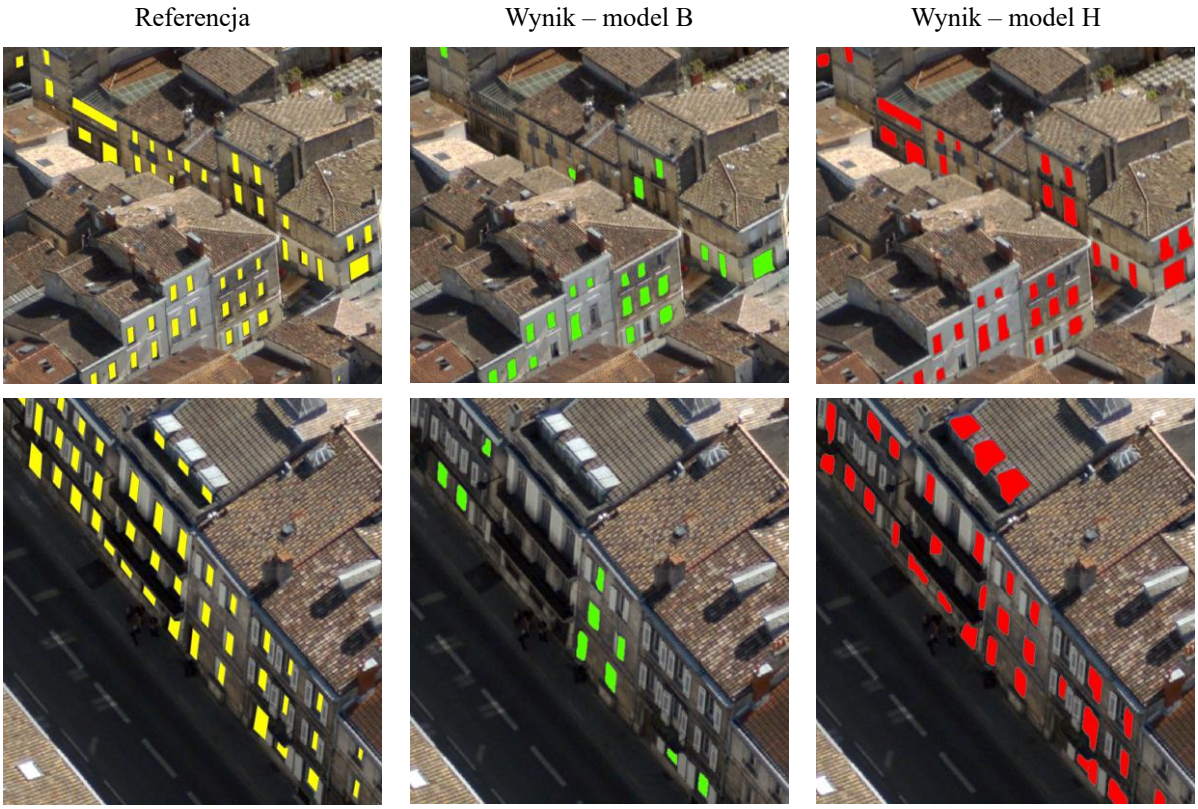
Budynek	Liczba okien	Liczba zdjęć w pełnej rozdzielczości
Wydział EITI	936	30
Wydział Transportu	266	30

Podobnie jak dla danych z Francji, zdjęcia o rozdzielczości 14192 x 10640 piksela zostały podzielone na kafelki, aby możliwe było przeprocesowanie. Ponadto zobrazowania lotnicze charakteryzuje różna wartość terenowego rozmiaru piksela (GSD), w przypadku Bordeaux było to 5cm, a dla zdjęć z Warszawy 2.5cm.

9.2. Wyniki modelu Segment Anything Model dla okien

Bordeaux

Pierwszym z analizowanych zestawów były dane z Bordeaux, dla których przetestowane zostały dwie wersje modelu - ViT-H i ViT-B. Pierwsze wyniki eksperymentów i wizualna ocena wykazały, że model ViT-H uzyskał dużo wyższe dokładności względem modelu o mniejszej liczbie parametrów - ViT-B, co oczywiście było dość spodziewanym rezultatem. Mimo, iż czas przetwarzania kafelka dla modelu ViT-H był prawie dwukrotnie dłuższy, przewaga w dokładności wyników była na tyle wysoka, że do dalszej analizy i porównania wyników w ocenie ilościowej zdecydowano się przenalizować wyniki z modelu SAM ViT-H. Porównanie wyników z różnych modeli przedstawia Rysunek 37.

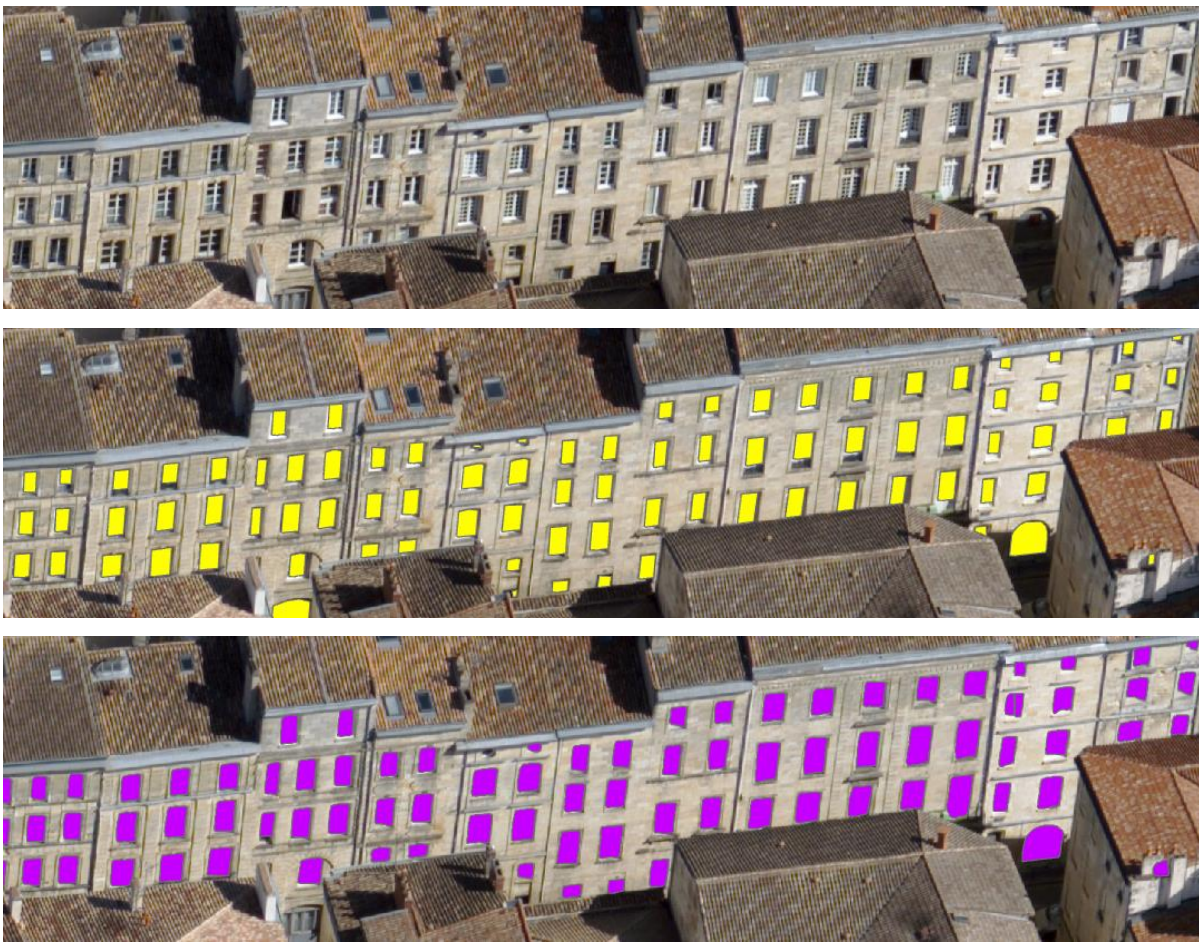


Rysunek 37 Porównanie wyników modelu SAM dla dwóch różnych wariantów (ViT-H i ViT-B) z danymi referencyjnymi.

Na podstawie analizy wizualnej można stwierdzić, że SAM wykazał zdolność segmentacji okien bez dotrenowania modelu. Największe problemy występowały w przypadku zacienionych obszarów oraz kiedy okna były odwzorowane na obrazie pod zbyt dużym kątem. Zdarzało się również tak, że segmenty nie odzwierciedlały poprawnie granic, łącząc w jeden obiekt np. okno i okiennice. W niektórych przypadkach mimo poprawnej segmentacji maska nie obejmowała całego obiektu. Jednakże jak widać na przykładowych obrazkach SAM generuje dość dobrze granice okien, co daje przewagę rozwiązania segmentacji nad detekcją. Przykładowe wyniki przedstawione zostały poniżej (Rysunek 38, Rysunek 39, Rysunek 40, Rysunek 41).

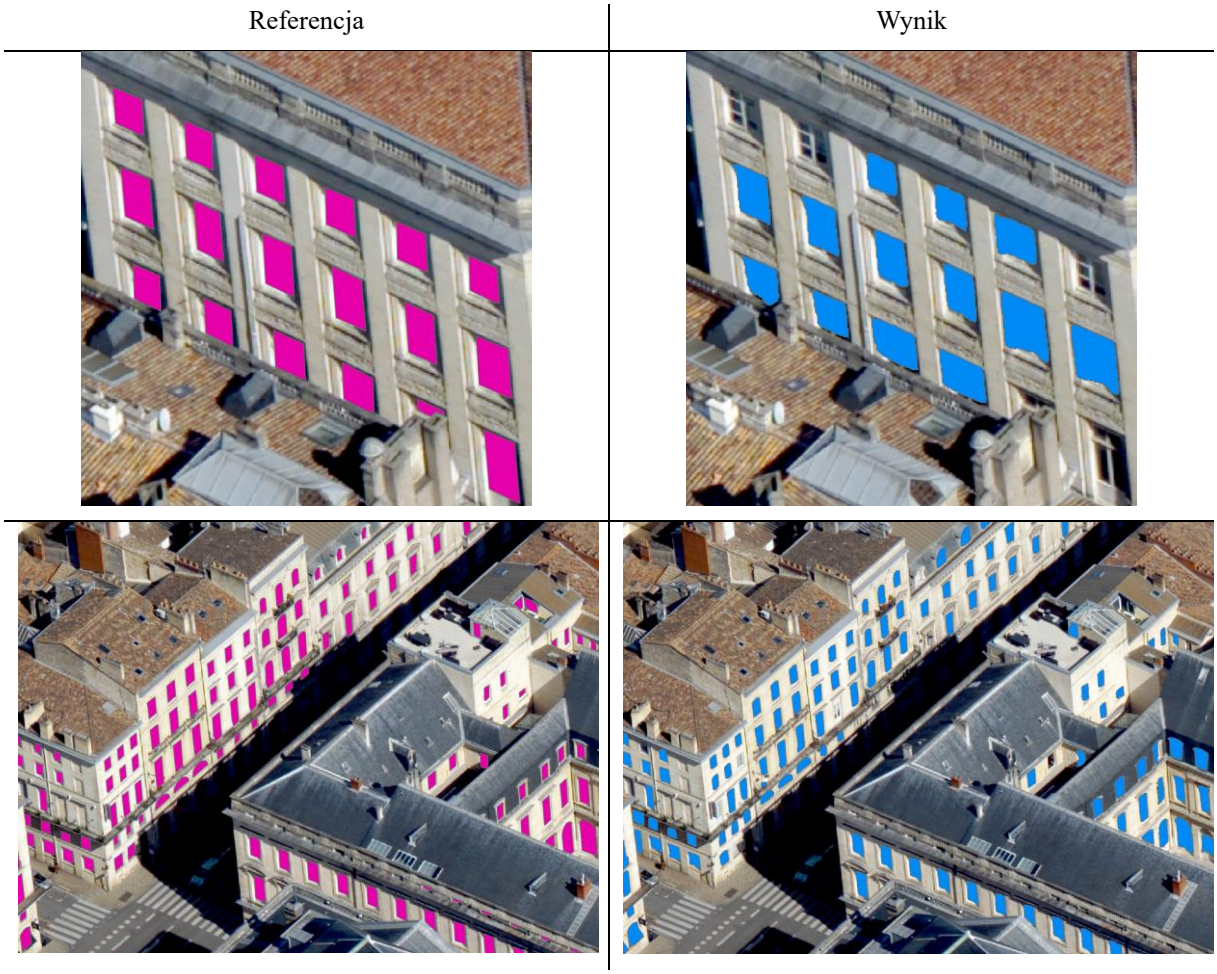
a)  b) 

Rysunek 38 Wynik modelu SAM dla segmentacji okien na ukośnych zdjęciach lotniczych z Bordeaux (a) dane referencyjne, (b) wynikowe segmenty.



Rysunek 39 Wynik modelu SAM. Od góry: obraz źródłowy, referencja, wynik.

Rysunek 40 Wynik modelu SAM. Od lewej: obraz źródłowy, referencja, wynik.



Rysunek 41 Porównanie wyniku modelu SAM względem referencji.

Metryki do oceny dokładności, które wykorzystano do analizy wyników to indeks Jaccarda, wartość precyzji oraz dokładność wyrażone w procentach. Ponadto jako prawidłowy segment był brany pod uwagę tylko ten, dla którego wartość IoU (Intersection over Union) była wyższa niż 0.5. Oprócz podstawowych metryk zestawiono również średnią wartość IoU, błąd lokalizacji (położenia) wyrażony w pikselach i spójność pikseli. Poniżej przedstawiono

metryki: (3) Indeks Jaccarda, (4) Precyzja, (5) Dokładność, które zastosowano na potrzeby badań:

$$Indeks\ Jaccarda = \frac{TP}{(TP + FP + FN)} \quad (3)$$

$$Precyzja\ [\%] = \frac{TP}{(TP + FP)} * 100\% \quad (4)$$

$$Dokładność\ [\%] = \frac{TP}{(TP + FN)} * 100\% \quad (5)$$

Tabela 21 Wyniki dla modelu SAM względem danych referencyjnych – obszar Bordeaux.

	MODEL ViT-B SAM	MODEL ViT-H SAM
Ground truth labels	12816	12816
Predicted labels	8440	10440
True Positives	6146	7712
False Positives	2276	2768
False Negatives	4349	2179
IoU	**0,714**	**0,745**
Jaccard	**0,497**	**0,628**
Pixel identity	**0,981**	**0,987**
Localization error	**8,142**	**8,025**
Accuracy [%]	**58,561**	**77,970**
Precision [%]	**72,976**	**73,588**

Porównując wyniki segmentacji do danych referencyjnych można stwierdzić, że lepsze dokładności uzyskano dla modelu ViT-H, co oczywiście potwierdziło analizę wizualną. Osiągnięta dokładność 78% dla modelu SAM (ViT-H) świadczy o potencjale wykorzystania podejścia. Należy podkreślić, że model nie był w żaden sposób douczany. Indeks Jaccarda również był znacząco wyższy dla modelu ViT-H. Wartości pozostałych wskaźników dla obu modeli ukształtowały się na podobnym poziomie. Parametrem, na który warto zwrócić uwagę to błąd lokalizacji, który w obu przypadkach wyniósł około 8 pikseli.

Powyższe porównanie zostało wykonane na poziomie masek (pikseli). Należałoby się jednak zastanowić jakie podejście do porównywania wyników segmentacji w tym przypadku byłoby

najlepsze oraz jaka dokładność segmentacji jest wystarczająca. Czy błąd położenia, który wynosi np. 5 pikseli powinien być wartością graniczną czy powinna zostać uwzględniona jeszcze rozdzielczość zdjęć?

Obiekty na zdjęciach mogą być odwzorowane z różnych perspektyw – może być to widok zbliżony do ortogonalnego, bądź widok ukośny, gdy elementy odwzorowane są pod pewnym kątem, co wpływa na to ile pikseli przypada na dany obiekt. W przypadku gdy fasada jest w ujęciu frontalnym, proporcje okien są zachowane, co ułatwia pomiar. Widok pod kątem powoduje, że okna wydają się „ściśnięte na obrazie", a co za tym idzie prowadzi to do zmiany pozornego rozmiaru obiektu, wpływając na liczbę reprezentujących go pikseli.

Oprócz wpływu perspektywy, zniekształceń i kąta pod jakim obiekt jest widoczny na zdjęciu należy wziąć pod uwagę również niejednorodność rozmiarów pikseli oraz zmienną skalę w obrębie zdjęcia podczas oceny wyników segmentacji okien na obrazach ukośnych. Problemy związane ze zmianą rozdzielczości terenowej (ang. *Ground Sample Distance* – GSD) w obrębie zdjęcia ukośnego dokładnie opisał Wojciech Ostrowski w swojej rozprawie doktorskiej.

Dlatego też różnice te mogą powodować, że metryki oceny, takie jak IoU będą zaniżane lub zawyżane. Nawet dobrze posegmentowane zdjęcie i wyodrębnienie okna na obrazie ukośnym może nie pasować idealnie do kształtu okna referencyjnego, zmniejszając dokładność na poziomie pikseli. Temat ten jest obszarem do wykonania analiz, jednak na potrzeby eksperymentów w niniejszej pracy zdecydowano się na klasyczne porównanie nakładania się masek na poziomie pikseli, a problem jedynie został nakreślony jako kierunek badawczy.

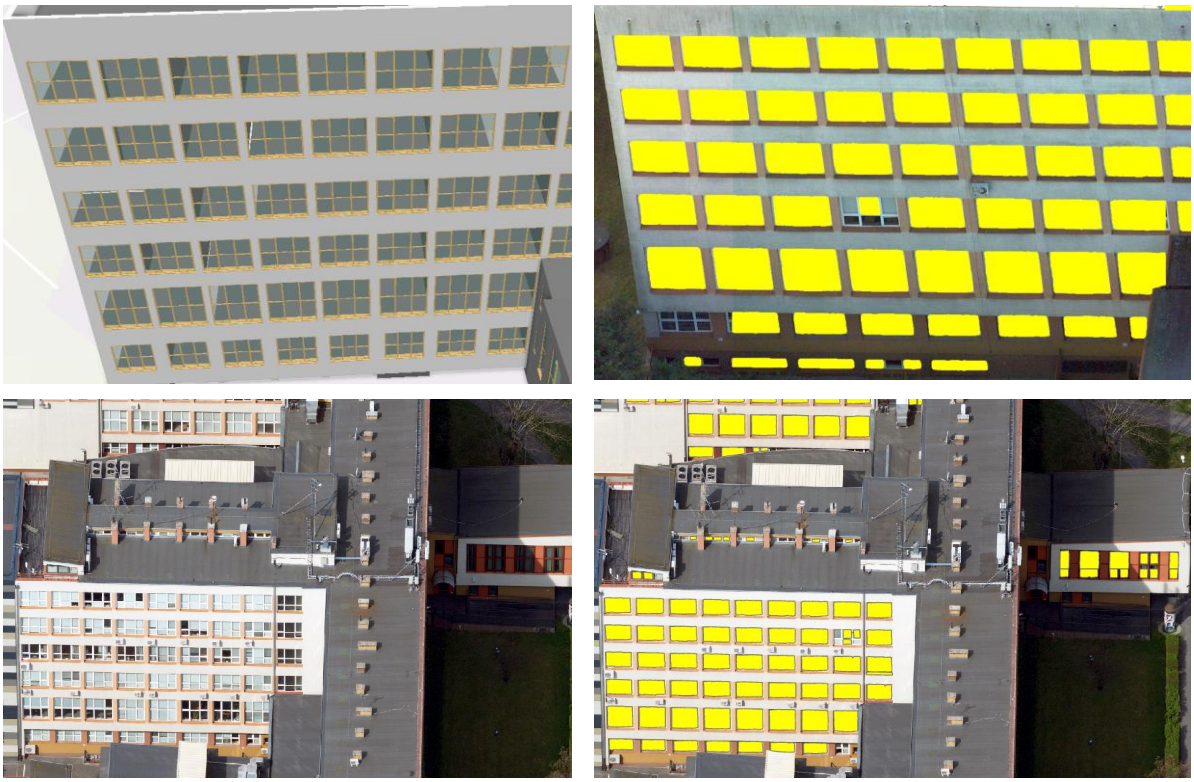
Biorąc pod uwagę fakt, iż model jest bardzo łatwy w użyciu, a oszczędność czasu obliczeniowego dzięki uniknięciu szkolenia od początku jest znacząca można powiedzieć, że takie rozwiązanie sprawdzi się w niektórych zadaniach. W przypadku konieczności utworzenia kompletnej bazy obiektów może stanowić rozwiązanie, które częściowo automatyzuje i przyspiesza cały proces. Oczywiście istnieje miejsce na poprawę możliwości uogólnienia modelu pod konkretne zastosowanie.

Warszawa

Na podstawie wyników z Bordeaux, gdzie znacząco lepszy okazał się model ViT-H, zdecydowano, że dla danych z Warszawy eksperymenty zostaną ograniczone do przetestowania tego modelu bez porównania z ViT-B, jak było to w przypadku poprzedniego obszaru. Poniżej

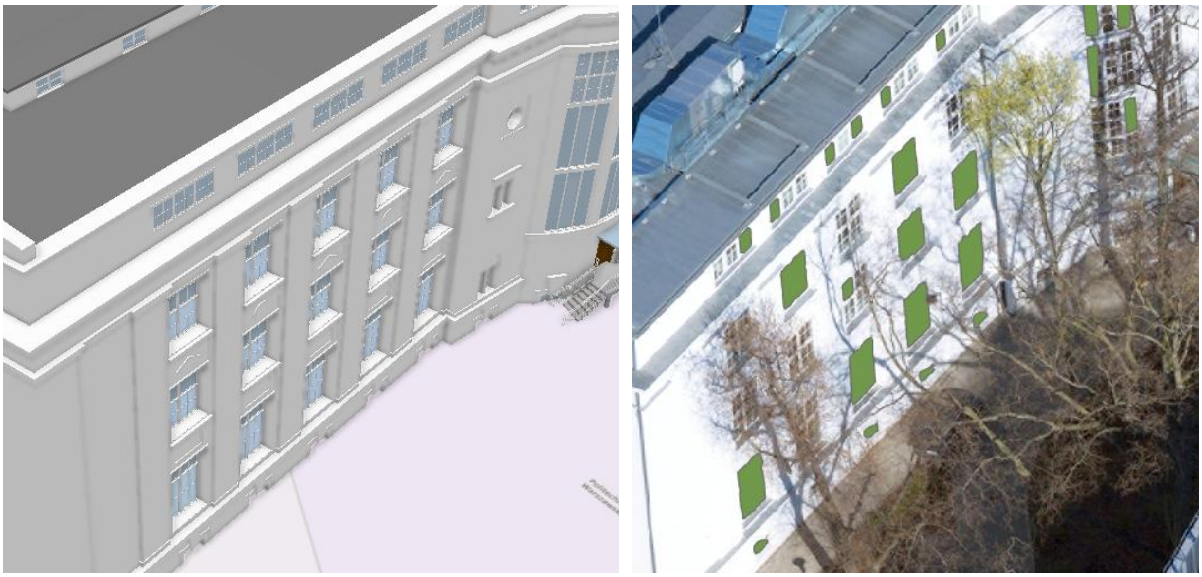
przedstawione zostały przykładowe wyniki segmentacji w zestawieniu wraz z danymi referencyjnymi (Rysunek 42, Rysunek 43).

(a) (b)



Rysunek 42 Wynik modelu SAM dla segmentacji okien na ukośnych zdjęciach lotniczych z Warszawy – Wydział EITI (a) dane referencyjne, (b) wynikowe segmenty.

(a) (b)



Rysunek 43 Wynik modelu SAM dla segmentacji okien na ukośnych zdjęciach lotniczych z Warszawy – Wydział Transportu (a) dane referencyjne, (b) wynikowe segmenty.

Na podstawie analizy wizualnej wywnioskowano, że znacznie lepiej SAM poradził sobie z segmentacją okien dla wydziału EITI. Duży wpływ w przypadku wydziału Transportu miały drzewa przysłaniające fasadę budynku na zdjęciach. Mimo wyboru zdjęć z kwietnia, kiedy jeszcze wegetacja nie była w pełni rozwinięta, gałęzie przysłaniające obiekt negatywnie wpłynęły na działanie modelu. Co więcej, należy zaznaczyć, że ludzkie oko jest w stanie poprawnie zinterpretować obraz i zidentyfikować okna, nawet jeśli jego część jest przysłonięta przez inny obiekt, z czym nie poradził sobie w tym przypadku model segmentacji. Prawdopodobnie dotrenowanie modelu i „pokazanie" kilku takich przypadków w zestawie treningowym byłoby dobrym rozwiązaniem na poprawę wyników.

Problem, który nie występował w przypadku danych z Bordeaux, a widoczny jest w przypadku Warszawy, to segmentacja szyb i wyodrębnianie poszczególnych skrzydeł okien przez model SAM.

Ze względu na to, że na wyrenderowanych obrazach referencyjnych okna potraktowane zostały jako segmenty w całości, bez podziału na poszczególne elementy zdecydowano się na „post-processing" danych i wykorzystanie filtrów morfologicznych w celu połączenia wyodrębnionych elementów okna w jeden obiekt.

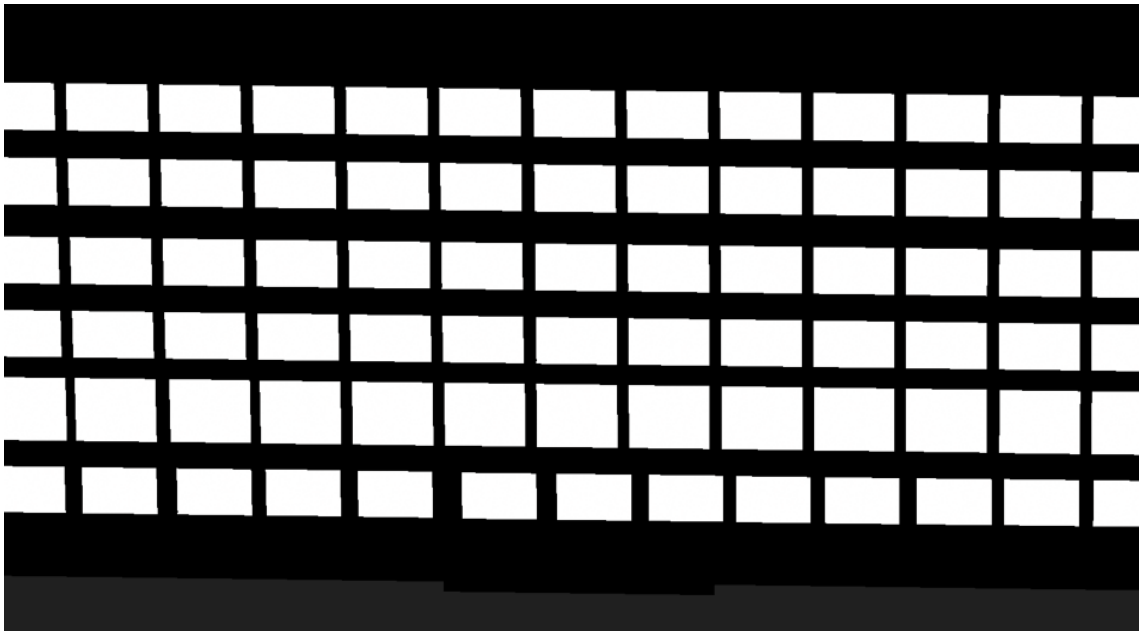
Przetwarzanie obrazów z wykorzystaniem operacji morfologicznych ma na celu uzyskanie różnych efektów, m.in. usunięcie szumów, zamknięcie dziur, dzięki czemu jest to często wykorzystywane narzędzie w analizie obrazu.

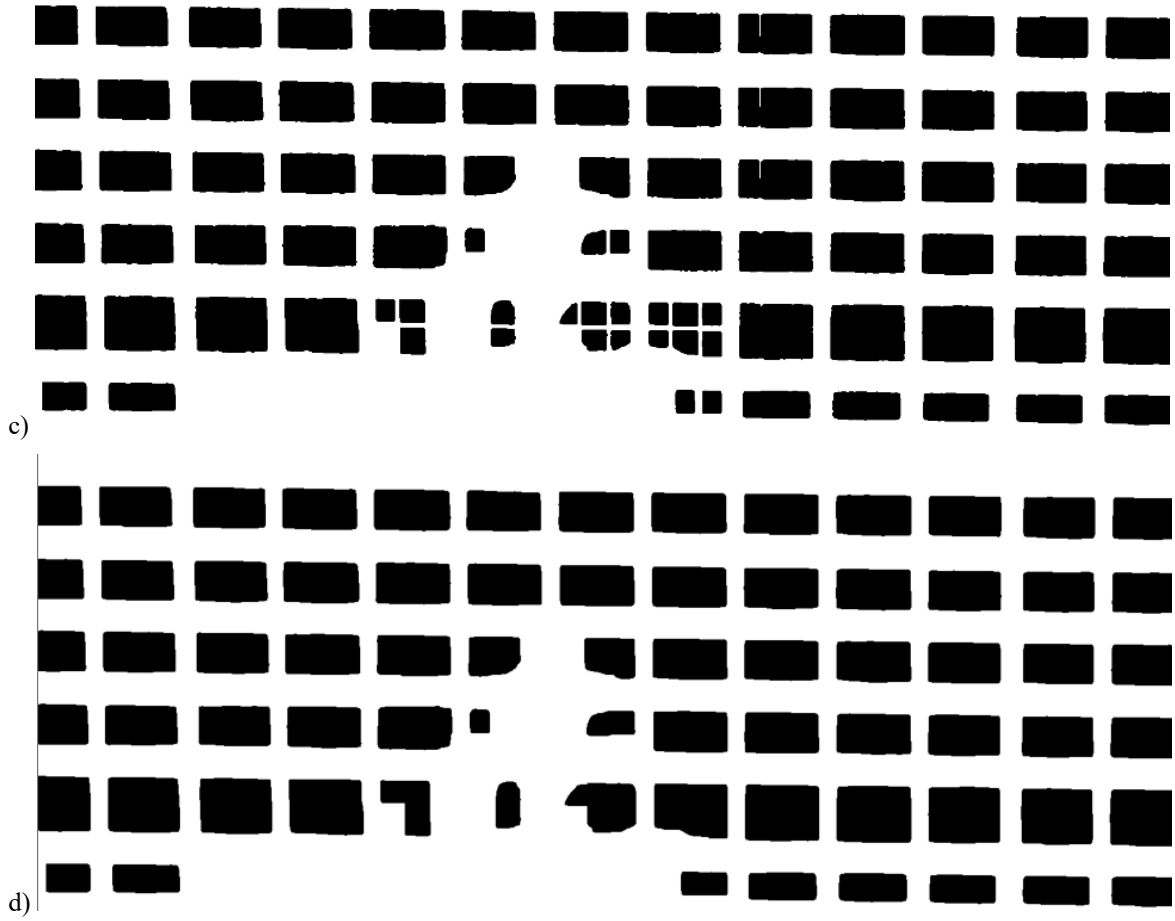
Na potrzeby badań zdecydowano się na wykorzystanie zamknięcia, które polega na następujących po sobie dylatacji i erozji. Dylatacja rozszerza granice, wypełniając małe dziury i łączy bliskie elementy, natomiast erozja powoduje zmniejszenie granic, dzięki czemu zmniejszony jest rozmiar obiektu. Połączenie to spowodowało wypełnienie luk w obiektach przy jednoczesnym zachowaniu ogólnego kształtu i rozmiaru. Przetestowane zostały różne rozmiary elementu strukturalnego (*ang. kernel*). Ostatecznie zastosowano rozmiar 15x15 dla wszystkich obrazów.

Operacja ta miała na celu zamknięcie dziur między obszarami i połączenia masek poszczególnych szyb w jeden wspólny segment okna. Zasada działania została pokazana poniżej (Rysunek 44).

a)

b)





Rysunek 44 Efekt działania operacji zamknięcia. A) zdjęcie pokazujące obiekt opracowania, B) dane referencyjne, C) wynikowe segmenty okien z modelu SAM, d) efekt działania morfologii w celu połączenia poszczególnych elementów okien w całość.

Analiza statystyczna pozwoliła na dokładną ocenę segmentacji okien. Ze względu na dużo słabsze wyniki zdecydowano się na bardziej rozbudowaną ocenę ilościową segmentacji okien dla Warszawy. Po pierwsze dla każdego zdjęcia o pełnej rozdzielczości policzona została referencyjna liczba okien oraz suma okien, które zostały wysegmentowane przez model SAM. Jak widać w tabeli poniżej (Tabela 22) w przypadku budynku Wydziału EITI uzyskano znacznie lepsze wyniki, co tylko potwierdziło wcześniejszą ocenę wizualną. Średnia (arytmetyczna) dokładność wyniosła dla budynku wydziału EITI 92,03%, a dla budynku wydziału Transportu 62,57%. Obliczono również wartość średnią kwadratową (*ang. root mean square*), która dla wydziału EITI ukształtowała się na poziomie 92,15%, a dla wydziału Transportu 63,73%.

Tabela 22 Zestawienie liczby okien dla każdego zdjęcia z uwzględnieniem danych referencyjnych i przewidywań modelu SAM dla obu analizowanych budynków oraz dokładności wyrażone w %.

WYDZIAŁ EITI				WYDZIAŁ TRANSPORTU			
Kierunek	Liczba okien (referencja)	Liczba okien (predykcja)	Dokładność %	Kierunek	Liczba okien (referencja)	Liczba okien (predykcja)	Dokładność %
L	210	187	89,05	R	110	53	48,18
L	212	193	91,04	R	119	54	45,38
R	243	208	85,60	R	125	63	50,40
R	242	202	83,47	R	122	58	47,54
L	243	214	88,07	R	126	63	50,00
L	243	223	91,77	R	117	59	50,43
F	273	261	95,60	L	113	59	52,21
F	272	262	96,32	L	125	67	53,60
F	274	262	95,62	L	122	67	54,92
B	137	130	94,89	L	125	63	50,40
B	135	122	90,37	L	115	60	52,17
B	135	128	94,81	L	125	53	**42,40**
F	135	128	94,81	F	113	75	66,37
F	135	131	**97,04**	F	113	79	69,91
F	136	130	95,59	F	131	101	77,10
B	272	260	95,59	F	134	87	64,93
B	272	257	94,49	F	111	80	72,07
B	289	263	91,00	F	108	80	74,07
F	271	260	95,94	B	134	94	70,15
F	272	261	95,96	B	132	103	78,03
B	134	127	94,78	B	116	79	68,10
L	245	215	87,76	B	113	76	67,26
L	241	201	**83,40**	B	133	101	75,94
R	241	216	89,63	B	131	103	78,63
R	240	220	91,67	L	131	76	58,02
L	240	228	95,00	L	131	66	50,38
R	238	221	92,86	L	130	101	77,69
R	241	213	88,38	L	130	100	76,92
R	250	216	86,40	R	130	97	74,62
B	134	126	94,03	R	130	103	**79,23**

Na podstawie wykonanych eksperymentów można powiedzieć, że model SAM poradził sobie z wykrywaniem okien na wysokorozdzielczych lotniczych zdjęciach ukośnych bez treningu na zbiorze uczącym. Uzyskane dokładności dla danych z Bordeaux ukształtowały się

na poziomie ok. 78%. Największe problemy w segmentacji występowały w obszarach zacienienia i w przypadku mniejszych okien. Dla drugiego obszaru badawczego (Warszawy) wyniki dla budynku wydziału EITI były znacznie lepsze – ok. 92%. Znacznie gorzej SAM poradził sobie z segmentacją wydziału Transportu, osiągając dokładność ok. 63%.

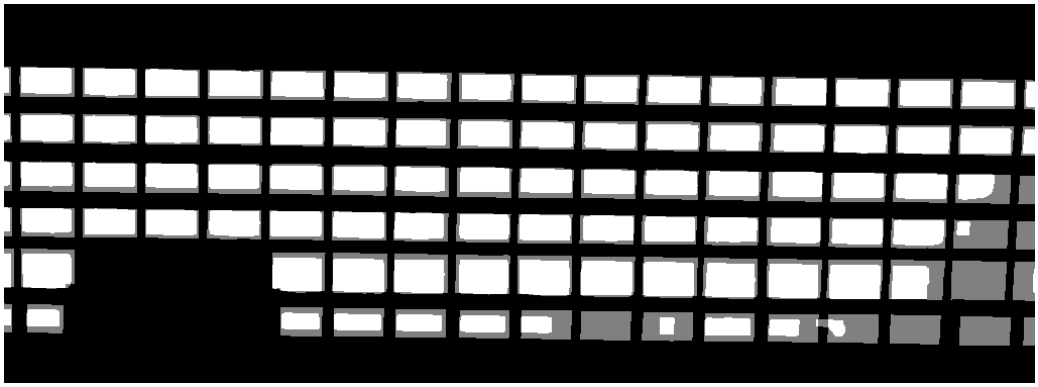
Biorąc pod uwagę duży wpływ drzew zasłaniających fasadę budynku w przypadku Wydziału Transportu zdecydowano, że statystyki pikselowe segmentacji okien zostaną wykonane tylko dla budynku Wydziału EITI. Metryki do oceny dokładności zastosowano analogiczne jak dla pierwszego obszaru badawczego – Bordeaux.

Tabela 23 Wyniki dla modelu SAM względem danych referencyjnych – obszar Warszawa, budynek Wydziału EITI.

Kierunek	IoU	Jaccard Index	Pixel identity	Precision [%]	Accuracy [%]
L	0,770	0,801	0,968	92,51	85,64
L	0,750	0,863	0,964	94,79	90,55
R	0,747	0,775	0,966	91,63	83,41
R	0,746	0,794	0,951	94,98	82,90
L	0,718	0,828	0,956	95,28	86,33
L	0,714	0,796	0,953	90,46	86,90
F	0,768	0,956	0,903	99,62	95,60
F	0,757	0,942	0,893	98,47	95,56
F	0,746	0,935	0,885	98,47	94,85
B	0,645	0,920	0,958	99,21	91,97
B	0,652	0,917	0,961	99,18	92,37
B	0,665	0,948	0,948	99,21	92,59
F	0,662	0,911	0,942	97,62	93,18
F	0,654	0,897	0,944	96,06	93,13
F	0,634	0,933	0,949	99,21	94,03
B	0,731	0,956	0,880	99,62	95,59
B	0,759	0,930	0,881	98,83	94,07
B	0,726	0,910	0,881	99,62	91,29
F	0,759	0,959	0,871	99,62	95,94
F	0,746	0,941	0,880	98,47	95,54
B	0,633	0,925	0,947	99,19	93,18
L	0,713	0,774	0,942	90,57	84,21
L	0,741	0,789	0,955	95,98	81,62
R	0,742	0,840	0,958	95,78	87,18
R	0,745	0,824	0,955	92,63	88,16
L	0,724	0,893	0,947	96,44	92,34
R	0,744	0,891	0,958	96,38	92,21
R	0,744	0,793	0,955	91,55	85,53
R	0,744	0,762	0,951	91,55	81,93
B	0,678	0,910	0,949	99,19	91,05

Jak widać powyżej (Tabela 23) przeliczone zostały wskaźniki: indeks Jaccarda, tożsamość pikseli (*pixel identity*), dokładność (*accuracy*) oraz precyzja (*precision*), co zapewniło wgląd w to, jak dobrze działa model na poziomie pikseli. W kontekście uzyskanych wyników jako pozytywną predykcję w analizie statystycznej uznawano okna, których wartość IoU względem referencji wyniosła więcej niż 0.40.

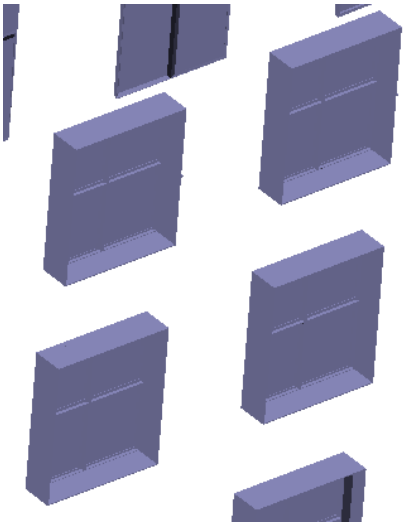
Warto zwrócić uwagę na wysokie wartości precyzji i dokładności, które dla większości zdjęć wyniosły ponad 90%. Wysoka wartość precyzji wskazuje na niewielką liczbę wyników fałszywie dodatnich. Warto jednak zaznaczyć, że jeśli model wykrył np. jedną szybę, ale IoU było niższe niż 0.40 względem referencji to było uznawane jako nieprawidłową predykcję. W tym aspekcie należałoby się dokładnie zastanowić jaka wartość powinna zostać przyjęta jeśli chodzi o nakładanie pikseli sklasyfikowanych jako okna z rzeczywistym stanem, aby uznać, że jest to poprawny wynik. Przykładowy wynik, gdzie okno zostało niewykryte w pełni pokazuje Rysunek 45. Na szaro zostały zaznaczone segmenty referencyjne, zaś białe to wynik z modelu. W tym przypadku niektóre okna podczas analizy statystycznej zostały uznane jako błędne. Powszechnie domyślną wartością progu IoU jest 0.50. W takim przypadku może niższy próg np. 0.30 byłby również akceptowalny, a poprawa wyników mogłaby nastąpić w przetwarzaniu końcowym. Zależy to oczywiście od wymagań odnośnie precyzji lokalizacji i kształtu obiektu, a także przeznaczenia w jakim celu mają być wykorzystywane wyniki.



Rysunek 45 Przykładowy wynik modelu SAM – porównanie wysegemntowanych okien na poziomie pikseli – szary kolor oznacza dane referencyjne, natomiast biały kolor to wynikowe okna z modelu SAM (budynek Wydziału EITI).

Pomimo wysokich wartości dokładności, IoU średnio wyniosło 0.72. Mimo iż wartość ta wydaje się niska jak na obszar pokrycia pomiędzy predykcją a maską referencyjną, to przyglądając się dokładniej wynikom należy zwrócić uwagę na to skąd taka różnica się bierze. Jak wspomniano w przypadku Warszawy odniesieniem była informacja o oknach

pozyskana z modeli BIM, gdzie okno było traktowane jako całość, uwzględniając wszystkie elementy, takie jak szyba, ramy, okiennice i inne powiązane z nim części (Rysunek 46).



Rysunek 46 Przykład okna z modelu referencyjnego BIM.

Segment referencyjny był zatem większy i zawierał więcej szczegółów, zaś model SAM identyfikował tylko szklaną część okna, skupiając się na mniejszym, bardziej specyficznym regionie bez dodatkowych elementów (Rysunek 47). Różnica w reprezentacji obiektu wyjaśnia niższe IoU oraz powód niedopasowania rozmiarów masek, a model nadal wykrywa istotną część obiektu. Dlatego też IoU na poziomie 0.72 można uznać za dobry, w szczególności gdy złożoność zadania nie wymaga szczegółowości obiektu na poziomie modelu BIM.

a) 

b) 

Rysunek 47 Różnica w reprezentacji okna – a) segmenty wynikowe z modelu SAM, b) segmenty referencyjne.

Modele uczenia głębokiego radzą sobie w zadaniach detekcji i segmentacji, osiągając dokładności powyżej 90-95%, a niewielkim nakładem pracy można uzyskać dobre wyniki, gdy mamy dostęp do danych uczących. Podejście, w którym bazuje się jedynie na segmentacji obrazu bez douczania modelu może być skuteczne w automatyzacji tworzenia zbioru uczącego.

Ogólnie rzecz biorąc, SAM wykazał duży potencjał w zakresie segmentacji ukośnych zdjęć lotniczych w celu wykrywania okien, a dalsze badania w tym obszarze mogą doprowadzić do jego powszechnego zastosowania do zadań z zakresu pozyskiwania geoinformacji.

Można by założyć rozwiązanie, w którym wykrywanie okien działa dwuetapowo, najpierw wykorzystuje się detektor do wstępnego wykrycia, a wynikowe bounding-boxy okien z detekcji mogłyby stanowić podpowiedź wejściową dla modelu SAM, który wygenerowałby dokładne segmenty z wyznaczonymi poprawnie granicami obiektów. Widać więc, że w niektórych specjalistycznych scenariuszach segmentacja może okazać się lepszym rozwiązaniem z przewagą nad detekcją.

10. Wyznaczanie lokalizacji obiektów z wykorzystaniem wyników uczenia maszynowego

Wiedza o położeniu wykrytego obiektu w układzie współrzędnych pikselowych zdjęcia może być wykorzystywana w różnych aplikacjach. Jednakże zaletą wykorzystania do wykrywania obiektów danych fotogrametrycznych jest fakt, że dane te zazwyczaj mają nadaną georeferencją. Można zatem powiedzieć, że pełne wykorzystanie potencjału lotniczych zdjęć ukośnych, to pójście krok dalej niż tylko wykrycie lokalizacji obiektu w przestrzeni obrazu, a dokładniej wyznaczenie położenia obiektu w terenowym układzie współrzędnych.

W przypadku zdjęć lotniczych, posiadając elementy orientacji wewnętrznej i zewnętrznej możliwe jest dokładne wyznaczenie współrzędnych terenowych wykrytego na obrazie obiektu, jednocześnie umożliwiając zautomatyzowanie procesu zasilania baz danych obiektów miejskich. Wykorzystanie wysokorozdzielczych lotniczych zdjęć ukośnych może zatem pozwolić zastąpić manualną inwentaryzację czy inspekcję terenową w celu zebrania danych o infrastrukturze i przestrzeni miejskiej wraz z odniesieniem przestrzennym. Realizacja tego zadania może być wykonana na kilka różnych sposobów. W przypadku przeniesienia wyników detekcji ze zdjęć do państwowego układu odniesienia można wyróżnić dwa podstawowe scenariusze:

- Zrzutowanie wyników detekcji na powierzchnię odniesienia (model mesh 3D, NMPT),
- Fotogrametryczne wcięcie w przód (w tym wypadku obiekt musi być wykryty na co najmniej dwóch zdjęciach).

Wyzwania badawcze, które można zdefiniować w tej części można zatem podzielić na kilka obszarów tematycznych:

- Wyniki z modelu nie są odniesione do układu pikselowego do zdjęcia o pełnej rozdzielczości. Aby zdjęcia mogły „przejść” przez model konieczne było „pocięcie” obrazów na mniejsze kafelki. Należy zatem zadbać, by przy transformacji obliczenia były wykonywane w odniesieniu do współrzędnych oryginalnego zdjęcia o pełnej rozdzielczości, gdyż dla takiego zdjęcia zapisywane są elementy orientacji zewnętrznej.
- Błędy modelu w wykrywaniu obiektów będą powodowały dalsze niedokładności w przeliczaniu do terenowego układu odniesienia. Pojawia się zatem pytanie czy manualna weryfikacja wyników detekcji przed przenoszeniem do terenowego układu jest niezbędna, czy ewentualne błędy mogą być wyeliminowane w kolejnych krokach?

- W jaki sposób uśredniać wyniki detekcji tego samego obiektu z kilku kierunków? Metoda klasteryzacji wyników z różnych kierunków i ujednolicenie do jednego obiektu w bazie danych powinna zostać dobrana tak, by możliwe było odrzucenie wartości odstających.
- Kolejnym pytaniem jest to, na ilu zdjęciach powinien być obiekt wykryty, aby można było uznać, że został poprawnie zlokalizowany przez model?
- Która z metod wyznaczania lokalizacji obiektów wykaże wyższe dokładności względem danych referencyjnych?

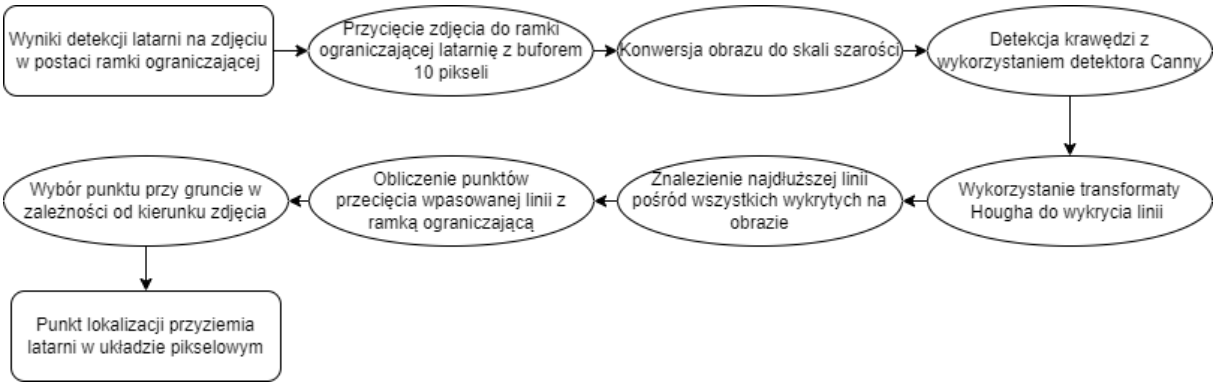
10.1. Metodyka wyznaczenia lokalizacji latarni

Detekcja latarni z wykorzystaniem modeli głębokiego uczenia może być wykonana z wysoką dokładnością, a ukośne zobrazowania lotnicze stanowią dobre źródło danych do tego typu zadania, co zostało wykazane we wcześniejszych eksperymentach.

Jeśli chodzi o obiekty małej architektury miejskiej takie, jak latarnie uliczne jednym z najważniejszych atrybutów w bazie danych jest położenie sytuacyjne (X,Y). Proponowana metodyka zakłada zatem przejście z układu pikselowego zdjęcia do układu terenowego uwzględniając wyniki detekcji w postaci ramek ograniczających, informacje o elementach orientacji wewnętrznej i zewnętrznej zdjęć oraz wykorzystanie nadliczbowości wykryć.

Pierwszym wyzwaniem badawczym w tej metodyce jest przejście z ramki ograniczającej latarnię do punktu przyziemia, który odpowiada lokalizacji. Zadanie to może być zrealizowane różnymi sposobami. Posiadając jednak wiedzę, że latarnia jest obiektem liniowym zdecydowano się na rozwiązanie oparte na wykrywaniu krawędzi na obrazie i zmianie wartości intensywności sąsiednich pikseli.

Po pierwsze wykorzystując wyniki detekcji - współrzędne ramki ograniczającej możliwe jest wykrycie w obszarze zainteresowania prostej, która definiuje słup latarni. Idąc krok dalej, możliwe jest znalezienie punktów przecięcia wpasowanej prostej z ramką ograniczającą. Ponadto dzięki temu, że znane jest położenie kamery w chwili wykonywania zdjęcia, możliwe jest wydobycie informacji o tym, w którą stronę „latarnia się kładzie na zdjęciu” dla każdego z czterech kierunków. W efekcie dla każdego kierunku możliwe jest określenie, który z dwóch punktów przecięcia prostej z ramką ograniczającą stanowi punkt przy gruncie. Można przedstawić ten proces na schemacie w następujący sposób (Rysunek 48).

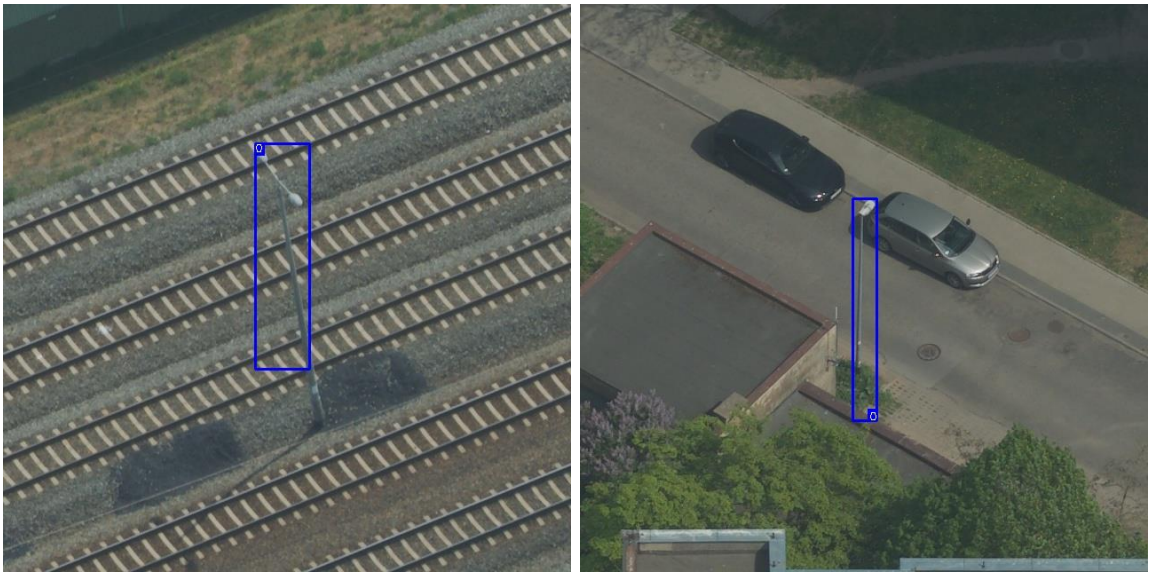


Rysunek 48 Schemat przedstawiający metodę wyznaczenia punktu oznaczającego przyziemie latarni na zdjęciu.

Działanie metody wyznaczenia punktu przyziemia pokazano dla przykładowej latarni (Rysunek 49). Przetwarzanie obrazu w tej części realizowane jest z wykorzystaniem biblioteki Python OpenCV. Zaczynając od lewej strony fragment obrazu przyciętego do wielkości ramki ograniczającej z buforem 10 pikseli konwertowany jest do skali szarości. Następnie z wykorzystaniem algorytmu wykrywania krawędzi Canny wykrywane są krawędzie na obrazie. Jeśli chodzi o szczegóły implementacyjne, na podstawie wykonanych testów ustawione zostały wartości progów intensywności pikseli - minimalna wartość parametru wyniosła 15, maksymalna natomiast 150. Wielkość maski Sobela wykorzystywanej do wykrywania pozostawiono domyślną wartość 3x3. Poszczególne kroki są niezbędne do wykonania przed wykrywaniem linii z użyciem transformacji Hougha. Oprócz obrazu, parametrami wejściowymi są *rho* i *theta*, za pomocą których możliwe jest określenie dokładności wykrywania linii. Na potrzeby eksperymentów przyjęto domyślną wartość 1 dla parametru *rho* oraz $\frac{\pi}{180}$ dla *theta*. Wynikowa prosta jest zdefiniowana za pomocą parametru stałej oraz nachylenia.

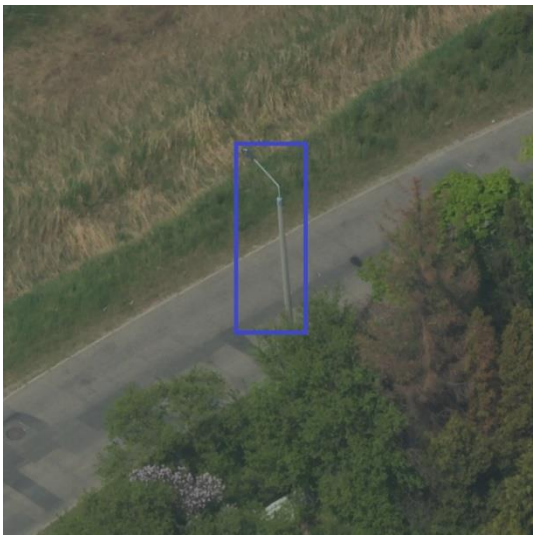

Rysunek 49 Przykład działania algorytmu wpasowania prostej i wyszukania punktu przyziemia latarni.

Analizując powyższe rozwiązanie oczywiście należy zwrócić uwagę na kilka aspektów, które mogą wpływać na wyniki. Naturalnie kluczową kwestią w wyznaczeniu punktu przyziemia latarni będzie dokładność uzyskana z detekcji obiektu z wykorzystaniem modelu głębokiego uczenia. Jeżeli ramka w pełni nie ogranicza latarni, będzie to miało wpływ na wyniki wyznaczenia punktu za pomocą wyżej opisanej metodyki. Przykłady detekcji latarni, w przypadku których występują pewne niedociągnięcia w dokładności ramki otaczającej obiekt pokazano poniżej (Rysunek 50).



Rysunek 50 Przykład wykrytych przez model latarni, dla których ramka ograniczająca nie otacza dokładnie obiektu.

Drugim z przykładów, o którym należy wspomnieć jest zasłanianie latarni przez inny obiekt terenowy na zdjęciu. W takim wypadku również algorytm nie zadziała w pełni poprawnie, gdyż lokalizacja przyziemia nie będzie określona w prawidłowy sposób. Mimo, iż model poradził sobie z detekcją latarni na poniższych przykładach (Rysunek 51), to określenie punktu na gruncie może stanowić problem.

a)  b) 

Rysunek 51 Przykłady latarni, które nie są w pełni widoczna na zdjęciu, gdyż przysłaniają je inne obiekty.

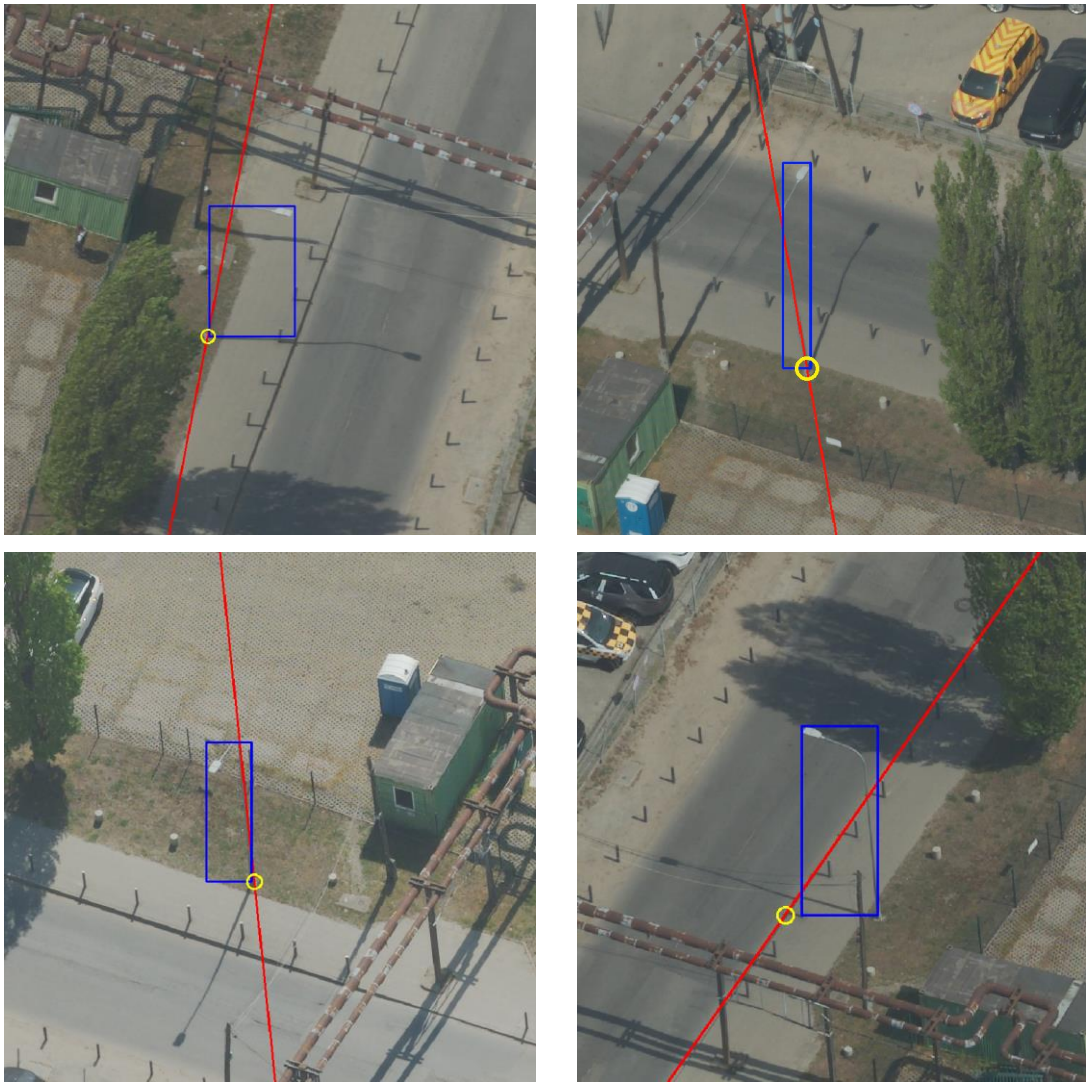
Jednakże naprzeciw takim przypadkom wychodzi nadliczbowość spowodowana dużymi pokryciami w bloku. Duża liczba zdjęć, na których widoczny jest dany obiekt oraz to, że jest odwzorowany z różnych kierunków pozwala na wyeliminowanie błędów i usunięcie wartości odstających w kolejnym kroku. Przykład tej samej latarni, która widoczna jest na Rysunek 51a) pokazano na zdjęciu z innego kierunku, gdzie możliwe jest określenie przyziemia latarni (Rysunek 52).

Rysunek 52 Przykład tej samej latarni, która widoczna jest na wcześniejszym rysunku 50a, ale odwzorowana na zdjęciu wykonanym z innego kierunku.

W tym miejscu należy również zaznaczyć, że w metodyce odrzucone zostały wykrycia na zdjęciach nadirowych. Zaproponowane podejście z wykrywaniem linii i punktu przyziemia nie sprawdziłoby się ze względu na różnice w odwzorowaniu tego typu obiektów między zdjęciami ukośnymi a pionowymi.

Oprócz widocznego wpływu wyników detekcji na kolejne kroki metodyki, obarczone błędem mogą być wyniki wpasowania linii z wykorzystaniem transformaty Hougha. Przykłady błędów, które najczęściej można było zidentyfikować podczas analizy wizualnej, to wpasowanie i wykrycie w obszarze ramki ograniczającej najdłuższej linii dla krawężnika, płotów, bądź innych obiektów liniowych podobnych do szukanych latarni. Zdarzały się również przypadki, w których żadna linia nie została wpasowana.

Poniższy przykład (Rysunek 53) przedstawia tę samą latarnię wykrytą na czterech zdjęciach z różnych kierunków, gdzie dla trzech metodyka zadziałała poprawnie, natomiast na jednym zdjęciu najdłuższa wpasowana linia odnosi się do krawężnika, który akurat odwzorował się w obszarze ramki otaczającej latarnię.



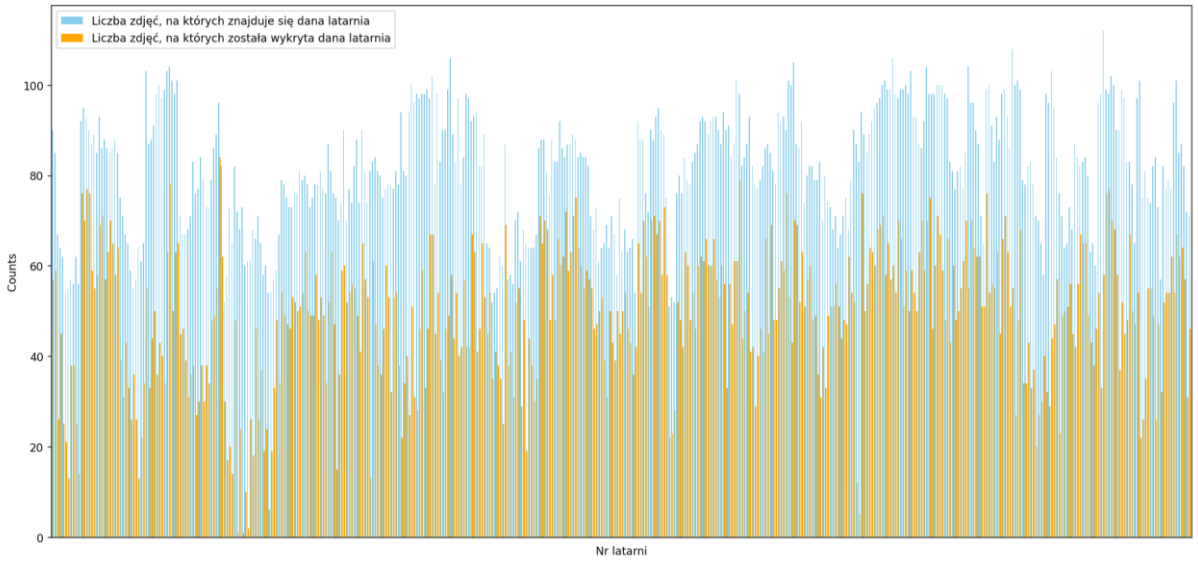
Rysunek 53 Wyniki definiowania przyziemia latarni na zdjęciu z wykorzystaniem wpasowania linii prostej za pomocą transformaty Hougha w obszarze ramki ograniczającej dla tej samej latarni odwzorowanej na zdjęciach z czterech różnych kierunków.

Mimo wspomnianych błędów, które mogą się zdarzyć, zauważalny potencjał podejścia wykazujący dobre wyniki wskazywania punktów przyziemia (Rysunek 54) oraz duża nadliczbowość pozwoliła na rozszerzenie metodyki i odrzucenie ewentualnych wartości odstających w kolejnym etapie związanym z łączeniem wyników z wielu kierunków.

Rysunek 54 Wyniki działania algorytmu definiowania punktów przyziemia na podstawie wykrytych za pomocą modelu ramek ograniczających latarnie.

Posiadając informacje o przyziemiu latarni ostatnim krokiem jest zatem transformacja współrzędnych do terenowego układu współrzędnych oraz zagregowanie wyników z wielu kierunków.

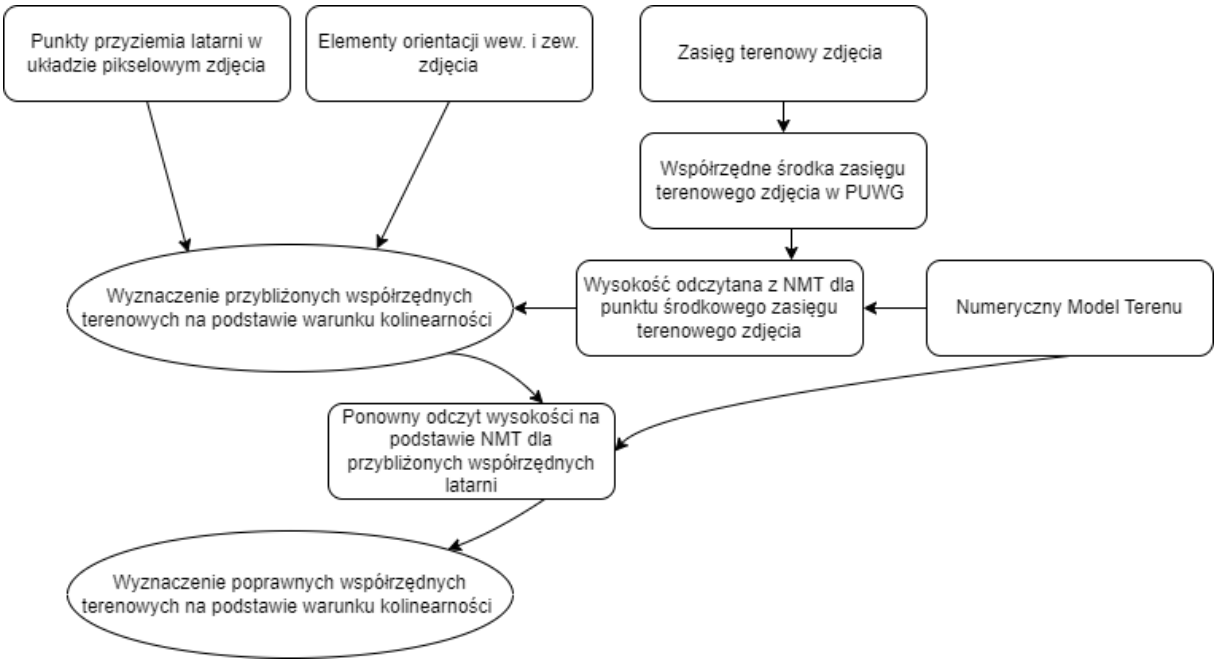
W tej części badawczej analizowane były wyniki detekcji 438 latarni pojedynczych i podwójnych. Dla każdej latarni ze zbioru obliczono statystyki wskazujące na ilu zdjęciach, na których obiekt został wykryty algorytm poradził z wpasowaniem prostej i znalezieniem punktu przyziemia. Średnio każda latarnia była widoczna na 81 zdjęciach, a punkt przyziemia średnio został znaleziony na 50 zdjęciach. Przykładowo dla jednej z latarni podwójnych, odwzorowanej na 86 zdjęciach, na 71 z nich algorytm wpasował prostą i znalazł punkt oznaczający dół. Poniższy wykres (Rysunek 55) przedstawia porównanie dla każdej latarni ze zbioru liczby zdjęć, na których się znajduje oraz liczbę zdjęć, na których znaleziony został potencjalny punkt przyziemia latarni. Analizując te statystyki można powiedzieć, że mimo iż rozwiązanie to pozwoliło na zdefiniowanie przyziemia latarni na około 60% zdjęć, to nadal można mówić o redundancji danych.



Rysunek 55 Porównanie liczby zdjęć, na których odwzorowana była latarnia oraz liczby zdjęć, na których algorytm poradził sobie z wpasowaniem prostej oznaczającej słup latarni i znalazł punkt przecięcia z ramką ograniczającą, w wyniku czego zdefiniowano punkty przyziemia każdego obiektu w układzie pikselowym zdjęcia.

Otrzymane w ten sposób współrzędne tłowe przyziemia latarni zostały przeliczone do współrzędnych terenowych na podstawie warunku kolinearności. Jako informacje wejściowe w tym etapie służyły współrzędne punktu przyziemia latarni wyrażone w pikselach, elementy orientacji wewnętrznej oraz zewnętrznej zdjęcia. W praktyce założenie to może być wykorzystane przy jednoczesnym posiadaniu informacji o wartości wysokości w terenowym układzie współrzędnych. Dlatego też zdecydowano się, że do obliczenia w pierwszej iteracji przybliżonych współrzędnych X,Y punktu wysokość zostanie odczytana z Numerycznego Modelu Terenu (NMT), który również często jest produktem pochodnym procesu fotogrametrycznego. Dla każdego ze zdjęć wyznaczone zostały zasięgi terenowe

(*ang. footprints*), a następnie wyznaczone zostały współrzędne środka każdego z równoległoboków. W drugiej iteracji ponownie odczytana była wartość z NMT, dzięki czemu możliwe było wyznaczenie poprawnych współrzędnych terenowych latarni. Proces został przedstawiony na poniższym schemacie (Rysunek 56).



Rysunek 56 Schemat przedstawiający realizację transformacji współrzędnych punktu z przestrzennego układu współrzędnych zdjęcia do terenowego układu współrzędnych na podstawie warunku kolinearności z wykorzystaniem Numerycznego Modelu Terenu i iteracyjnego podejścia do wyznaczania wysokości punktu.

W tym miejscu należy również zaznaczyć, że w procesie rzutowania punktu z przestrzennego układu współrzędnych zdjęcia do terenowego układu współrzędnych nie uwzględniono dystorsji obiektywu, gdyż w przypadku systemu Leica CityMapper mamy do czynienia ze zdjęciami pozbawionymi zniekształceń, które zostały wyeliminowane na etapie przetwarzania surowych zdjęć.

Otrzymana warstwa punktowa z wynikami rzutowania została poddana analizie wizualnej. Jako podkład służyła ortofotomapa. Jak widać poniżej (Rysunek 57) wyniki okazały się zadowalające i większość punktów poprawnie wskazywała położenie latarni. Oczywiście wcześniej wspomniane problemy wpłynęły na to, że nie wszystkie punkty poprawnie wskazywały lokalizację obiektu. Jednakże był to spodziewany efekt, ale dzięki dużej redundancji danych możliwe było wyeliminowanie wartości odstających i połączenie wyników w celu uzyskania najbardziej dokładnego położenia latarni w terenowym układzie odniesienia.

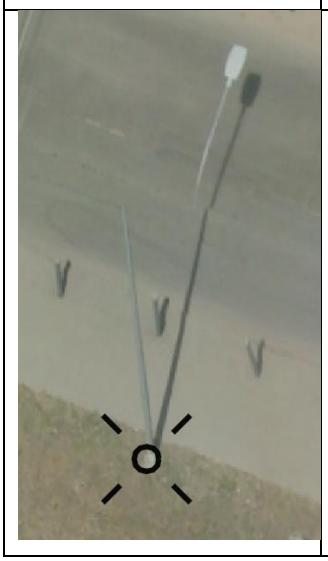
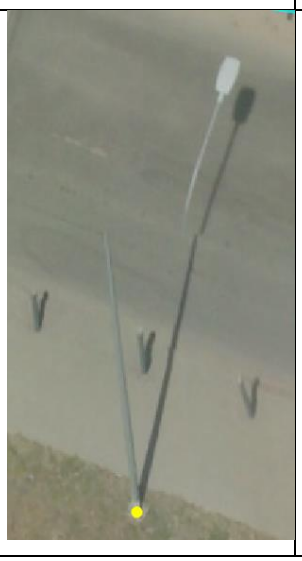
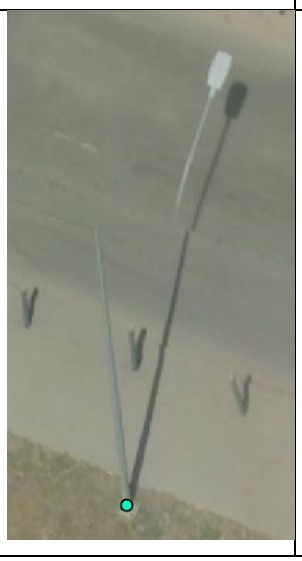
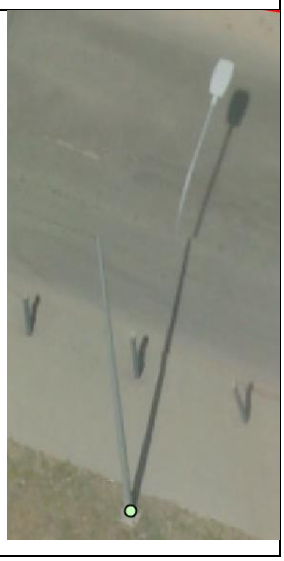
Rysunek 57 Wyniki rzutowania punktów z układu pikselowego zdjęcia do terenowego układu współrzędnych dla trzech przykładowych latarni. Jako podkład posłużyła ortofotomapa.

Analiza otrzymanych wyników pozwoliła na dobór techniki, która polega na grupowaniu danych, a mianowicie wykorzystano klasteryzację za pomocą algorytmu DBSCAN (*ang. Density-Based Spatial Clustering of Applications with Noise*) (Ester I in., 1996). Wykonanie klastrowania DBSCAN po pierwsze pozwoliło na pogrupowanie punktów, a następnie pozostawienie klastra, w którym znajdowała się największa liczba punktów. Umożliwiło to odrzucenie wartości odstających (Rysunek 58). Finalnie na podstawie pozostałych punktów wyznaczone zostały współrzędne środka ciężkości, które oznaczały położenie latarni w terenie.

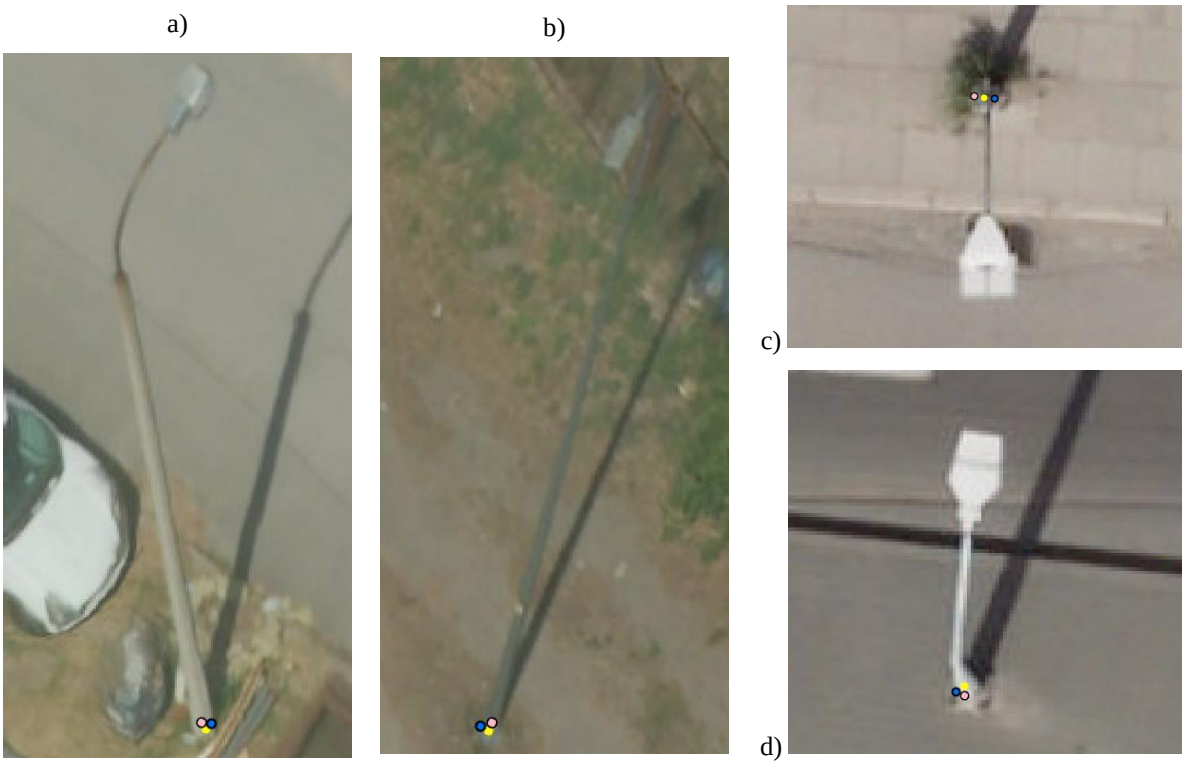
Rysunek 58 Przykładowe wyniki klasteryzacji DBSCAN i odrzucenie wartości odstających.

Drugim podejściem zastosowanym w ramach badań było wyznaczenie współrzędnych terenowych realizując fotogrametryczne wcięcie w przód. W tym przypadku jako dane wejściowe posłużyły tylko te wykrycia, które nie zostały wyeliminowane podczas klasteryzacji DBSCAN. Celem wykorzystania innego podejścia była weryfikacja czy wyznaczanie współrzędnych za pomocą wyrównania (triangulacji przestrzennej) znacząco poprawi wyniki względem danych referencyjnych. Wykorzystano do tego bibliotekę pigEOn rozwijaną w Zakładzie Fotogrametrii Teledetekcji i Systemów Informacji Przestrzennej na Wydziale Geodezji i Kartografii Politechniki Warszawskiej.

Przykładowe rezultaty przedstawiono poniżej (Rysunek 59, Rysunek 60). Jak widać wyniki uzyskane dla podejścia wykorzystującego warunek kolinearności i NMT do zrzutowania wykryć latarni do terenowego układu współrzędnych są zbliżone do danych referencyjnych. Ocena wizualna metody realizującej wcięcie w przód również pozwoliła potwierdzić, że obliczone pozycje latarni są poprawne. Oprócz analizy jakościowej dokonano również oceny ilościowej.

GESUT	REFERENCJA – manualne oznaczenie na ortofotomapie	Wynik z wcięcia w przód	Wynik rzutowania wykryć latarni na powierzchnię odniesienia
			

Rysunek 59 Porównanie wyników wyznaczenia położenia latarni ulicznej w terenowym układzie współrzędnych dla dwóch metod: fotogrametrycznego wcięcia w przód oraz rzutowania z wykorzystaniem NMT z danymi referencyjnymi wyznaczonymi na podstawie manualnego pomiaru na ortofotomapie i danymi GESUT (wykorzystana warstwa WMS z Geoportalu).



Rysunek 60 Wyniki wyznaczenia lokalizacji punktów dla latarni wyświetlone na podkładzie ortofotomapa (a i b) oraz na podkładzie true-ortofotomapa (c i d).Oznaczenia punktów: niebieski – wynik wcięcia w przód; różowy – wynik rzutowania; żółty – referencja.

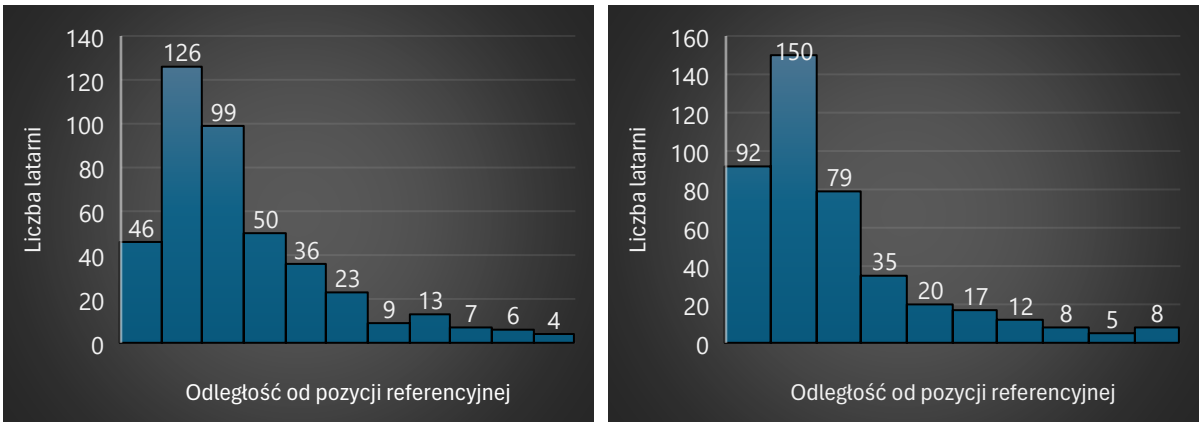
Liczba latarni poddanych analizie dla obu metod wyniosła 438. W pierwszym kroku przeliczone zostały odległości między lokalizacją punktu referencyjnego a położeniem wyznaczonym dwiema metodami. W przypadku rzutowania punktów spośród 438 latarni dla 19 z nich różnica odległości między punktem referencyjnym a wyznaczonym wyniosła powyżej 50cm. Dla 419 pozostałych latarni obliczona została wartość średnia, która ukształtowała się na poziomie 14,1cm. Druga metoda wyznaczenia położenia latarni za pomocą wcięcia w przód wykazała, że wartość średnia odległości między referencyjną lokalizacją latarni a wynikową z detekcji na zdjęciach lotniczych wyniosła 12,8cm. Ponadto dla 12 latarni odległość wyniosła powyżej 50cm. Aby jeszcze lepiej zdefiniować różnice w położeniu punktu wyznaczone zostały wartości RMSE (*ang. Root Mean Squared Error*) dla wszystkich latarni dla współrzędnych X i Y. Błąd średni kwadratowy odległości w obu metodach był większy dla współrzędnej X. Wyniki zestawione zostały poniżej w Tabela 24.

Tabela 24 Wyniki odległości między referencyjnymi lokalizacjami latarniami, a wyznaczonymi na podstawie wyników detekcji na lotniczych zdjęciach ukośnych i obliczeń fotogrametrycznych.

	RZUTOWANIE	WCIĘCIE
Razem wykrytych latarni: 438	419	426
powyżej 50cm	19	12
Wartość średnia odległości między referencyjną lokalizacją latarni a wynikową z detekcji na zdjęciach lotniczych [m]	0,141	0,128
RMSE X [m]	0,132	0,125
RMSE Y [m]	0,105	0,103

Na histogramach (Rysunek 61) uwzględnione zostały tylko latarnie, dla których odległość była mniejsza niż 50cm. Odległości między lokalizacją referencyjną a wyznaczoną z wykorzystaniem opisanej wyżej metodyki zestawiono na histogramach w przedziałach od 0 do 50cm w przedziałach 5cm.

(a) (b)



Rysunek 61 Rozkład odległości między referencyjną lokalizacją latarni, a wyznaczoną a) w procesie rzutowania oraz b) za pomocą wcięcia w przód. Przedziały podzielone co 5cm od 0 do 50cm.

Biorąc pod uwagę fakt, iż referencja w postaci punktów oznaczających lokalizację latarni była wykonywana manualnie poprzez fotointerpretację na podstawie ortofotomapy i true ortofotomapy o rozdzielczości 2,63 cm, a błąd fotointerpretacji wynosi ok. 2 pikseli można powiedzieć, że błąd położenia punktów referencyjnych wyniósł ok. 5cm. Uzyskane dokładności ukształtowały się na poziomie 13-14 cm względem danych referencyjnych, co daje zadowalający efekt. Ponadto dobór referencji, względem której były liczone błędy nie jest może najlepszym rozwiązaniem, jednak na takim poziomie dokładności należałoby odnieść się do danych pozyskanych tachimetrem bądź metodą GPS za pomocą odbiornika GNSS, co wiązałoby się z manualną pracą.

10.2. Metodyka wyznaczenia lokalizacji okien

Postępy technologiczne sprawiły, że fotogrametria bazująca na zdjęciach ukośnych stała się podstawową metodą rekonstrukcji 3D obszarów miejskich. Aktualnie pozyskiwanie bardzo szczegółowych, fotorealistycznych i wielkoskalowych fotogrametrycznych modeli mesh jest w pełni zautomatyzowane i wydajne (Gerke i in., 2016; Toschi i in., 2017). Jednakże w przypadku modeli siatkowych ograniczeniem wykorzystania w pełni potencjału tego typu produktów jest brak informacji o takich obiektach jak okna, dachy czy fasady. Wzbogacenie o ten element może ułatwić pełne wykorzystanie aplikacji geoprzestrzennych i skutecznie wdrożyć strategie planowania przestrzennego czy łagodzenia skutków katastrof (Wang i in., 2024).

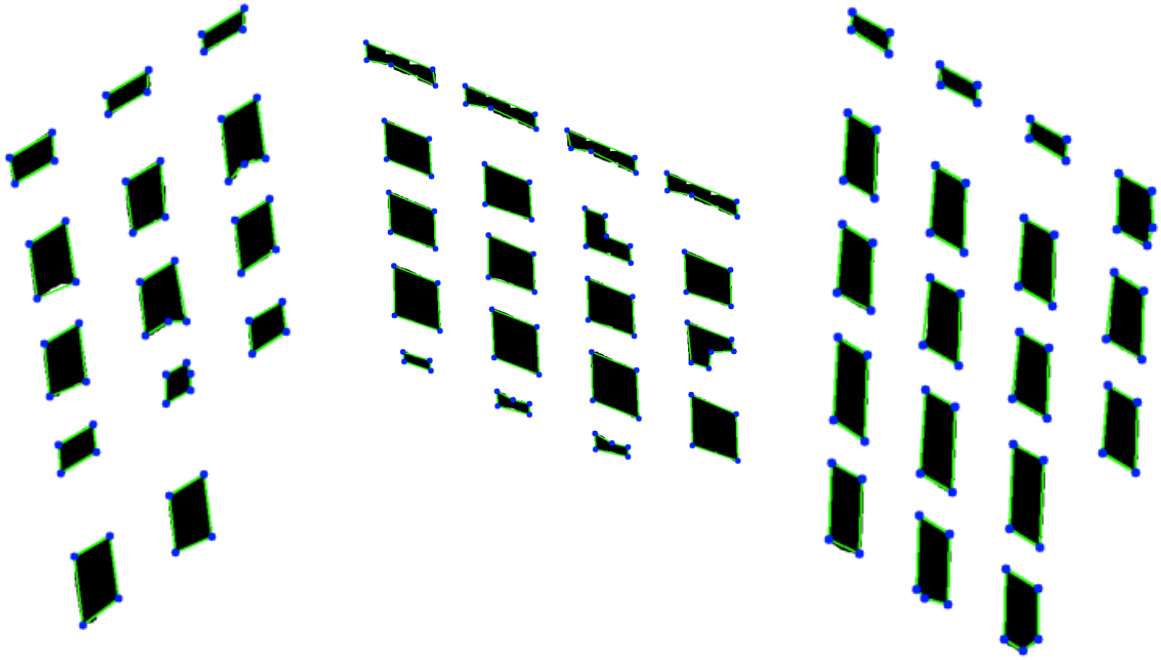
Przeniesienie informacji semantycznej otrzymanej w wyniku segmentacji wydaje się być większym wyzwaniem niż było to w przypadku obiektu takiego typu, jak latarnia. W kontekście obiektów takich jak okna ważnym aspektem jest połączenie informacji o lokalizacji okna na przykład w odniesieniu do modelu budynku, a nie tylko wyznaczenie lokalizacji punktu, jak w przypadku latarni ulicznej.

W procesie łączenia wyników segmentacji semantycznej przeprowadzonej na zdjęciach ukośnych można wyróżnić zatem kilka podejść. Posiadając model 3D budynku w postaci modelu mesh 3D lub modelu bryłowego możliwe jest przypisanie informacji semantycznej bezpośrednio do geometrii jako dodatkowego atrybutu. Jednym z rozwiązań proponowanych w literaturze jest zapisanie informacji semantycznej w postaci tekstury dla modeli budynków. Innym podejściem również wcześniej przytoczonym w przeglądzie literatury jest metodyka zaproponowana przez Mao i in., 2023. W swojej pracy posegmentowane okna ze zdjęć rzutują na model 3D w strukturze mesh, następnie udoskonalana jest struktura modelu, w tym wykorzystane jest wygładzenie siatki i ponowne mapowanie tekstury w celu rekonstrukcji i usunięcia deformacji.

W niniejszej pracy zdecydowano się zaś na dodanie informacji o oknach do modeli 3D w standardzie CityGML 2.0 na poziomie szczegółowości LoD2. Wynik segmentacji w ten sposób wnosi dodatkowe informacje do bazy danych o budynkach, dzięki czemu modele z dodatkowym atrybutem są wzbogacone o nowe dane, które mogą służyć do dalszych analiz przestrzennych.

W podejściu, gdzie wykorzystywane są zdjęcia istotna wydaje się strategia grupowania segmentów dla tego samego obiektu z różnych zdjęć i łącznia wyników z wielu kierunków.

W tym zadaniu można wyróżnić kilka sposobów jak rozwiązać problem. Jednakże jako, że wynikiem segmentacji były maski w postaci rastra, gdzie czarne piksele oznaczały okno, a białe reszta zdjęcia zdecydowano, że na podstawie analizy obrazu zostaną wyekstrahowane piksele oznaczające obiekt zainteresowania. Wykorzystując funkcje cyfrowego przetwarzania obrazów dostępne w bibliotece OpenCV możliwe było wyznaczanie współrzędnych pikseli narożników segmentów okien. Za pomocą metody Canny wykryto krawędzie w celu znalezienia konturów okien na obrazie i pozyskania informacji o narożnikach. Takie podejście pozwoliło na wyznaczenie punktów w układzie pikselowym, które reprezentowały narożniki okien i mogły zostać przeliczone do układu współrzędnych terenowych. Działanie tego kroku pokazano poniżej (Rysunek 62).



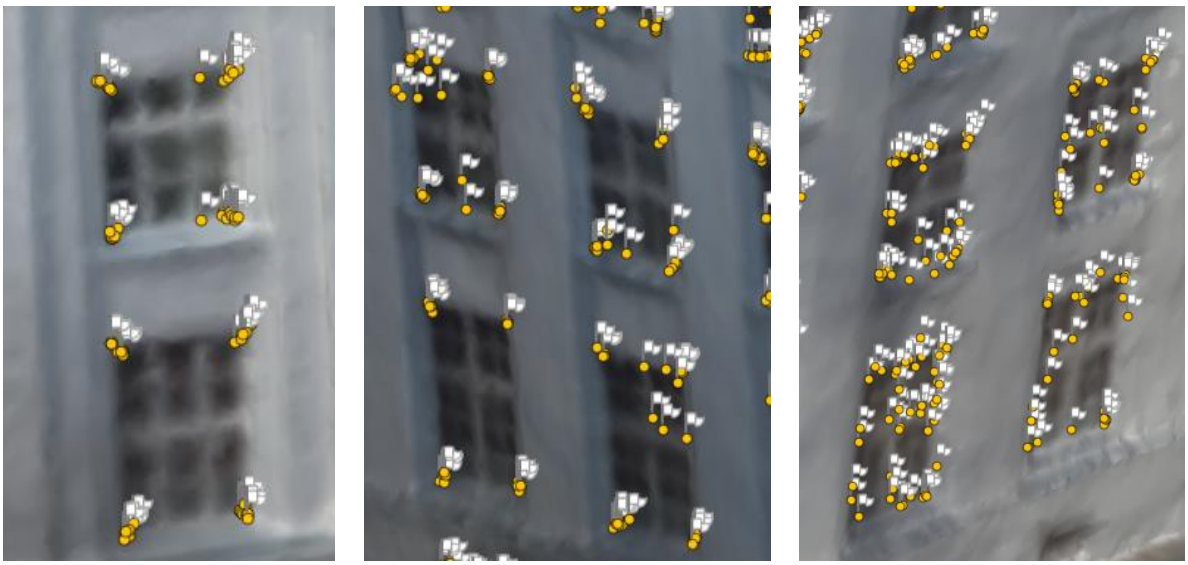
Rysunek 62 Ekstrakcja narożników segmentów okien w celu otrzymania współrzędnych pikselowych do dalszych obliczeń w układzie współrzędnych terenowych.

Współrzędne pikselowe znalezionych narożników okien zostały w automatyczny sposób zapisane w pliku tekstowym. W wyniku tego kroku nie było jeszcze informacji o lokalizacji okna w terenowych układzie, jednak ponownie wykorzystano matematyczne przeliczenia, aby uzyskać końcowy wynik, a ostatnim wyzwaniem było połączenie współrzędnych z wielu zdjęć i przypisanie informacji do modelu 3D. Tutaj pojawiło się również pytanie jak te informacje powinny być przechowywane w bazie danych obiektów topograficznych, aby wykorzystać potencjał informacyjny.

Kolejnym krokiem metodyki było wykorzystanie programu Agisoft Metashape Professional i Pythona do zautomatyzowania procesu wyznaczania współrzędnych terenowych 3D

narożników okien na podstawie współrzędnych pikselowych 2D wraz z odpowiadającymi im nazwami zdjęć zapisanymi w pliku *.csv. Lokalizacja wyznaczana była za pomocą funkcji „*chunk.model.pickPoint*”, która przekształca współrzędne obrazu 2D w punkt 3D na podstawie rzutowania promieni na powierzchnię odniesienia. Rzutowanie współrzędnych rastra może być wykonane zarówno na model mesh 3D, jak i na chmurę punktów czy też wykorzystanie modelu numerycznego – NMPT (numerycznego modelu pokrycia terenu) lub NMT (numerycznego modelu terenu). Jako, że model siatkowy jest jednym z produktów uzyskiwanych w procesie fotogrametrycznego przetwarzania zdjęć lotniczych zdecydowano się na wykorzystanie takiego rozwiązania. Ostatni krok to konwersja punktu 3D ze współrzędnych lokalnych na globalne przy użyciu macierzy transformacji (Rysunek 63).

W wyniku uzyskano dla obu budynków zestaw punktów oznaczających okna o współrzędnych X,Y,Z w państwowym układzie geodezyjnym. Ponadto fakt, iż budynek został odwzorowany na wielu zdjęciach pozwoliło na nadliczbowość wykryć zdjęć. Dało to również przewagę w momencie gdy okno nie zostało wysegmentowane na jakimś ze zdjęć, gdyż mogło zostać wykryte na innym zdjęciu z innego kierunku.



Rysunek 63 Przykład wyznaczenia lokalizacji okien wykrytych na zdjęciach. Wyniki segmentacji na wielu zdjęciach z różnych kierunków zrzutowane na model mesh 3D.

Tak więc po obliczeniu rzeczywistych współrzędnych okien, następnym krokiem było połączenie wyników w celu poprawy dokładności. Najprostszym podejściem byłoby wyznaczanie średniej z punktów dla każdego narożnika dla wszystkich okien. Jednakże patrząc na wyniki, gdzie w zasadzie przeliczone współrzędne wyświetlone na tle modelu to skupiska punktów bez informacji o tym, które to jest okno i jaki ma identyfikator, takie podejście nie byłoby prawdopodobnie optymalnym rozwiązaniem. Ponadto należy wziąć pod uwagę

na to, że mogą występować jakieś błędy, bądź różnice miedzy wynikami z różnych kierunków. Przykładowo gdy na sześciu zdjęciach okno zostało wysegmentowane poprawnie, a na trzech z innego kierunku częściowo było przysłonięte to współrzędne narożników będą się różniły.

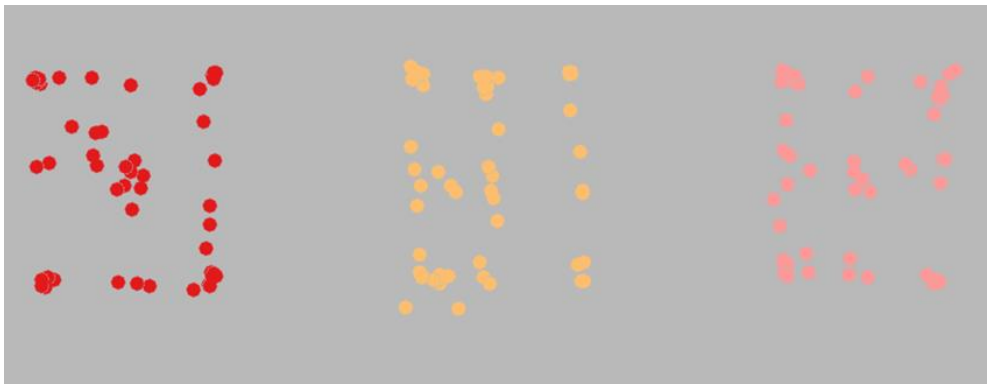
Inną propozycją mogłoby być uśrednienie wyników z nadaniem wag. Oznaczałoby to rozróżnienie pod względem tego, że większą wagę miałby te zdjęcia, które wykonane zostały z takiego kierunku, który powinien z lepszej perspektywy przedstawiać fasadę. Podejściem, które również prawdopodobnie wpłynęłoby pozytywnie na wyniki wykluczenie wartości odstających np. z użyciem technik takich jak RANSAC (*ang. random sample consensus*).

Ujednolicenie wyników z wielu kierunków w niniejszej metodyce zakłada klasyczne podejście. Etap ten można podzielić na kilka kroków:

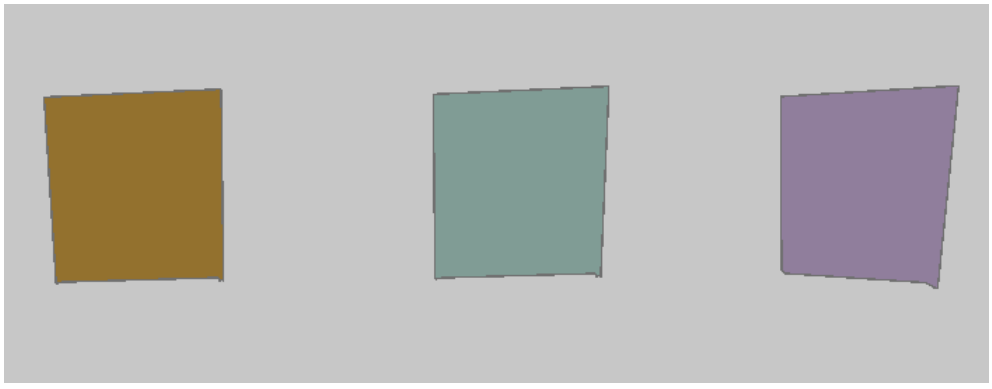
- odrzucenie wartości odstających z warstwy punktowej stanowiącej narożniki okien,
- klasteryzacja punktów w grupy oznaczających poszczególne okno,
- wykorzystanie narzędzi ArcGIS Pro w celu utworzenia wielokątów, które reprezentują określoną geometrię otaczającą grupę punktów wejściowych.

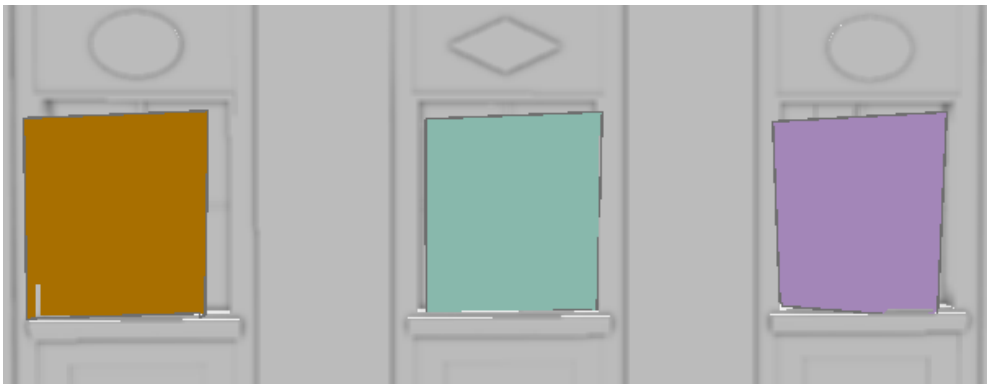
Poniżej pokazano dokładny przykład dla trzech okien, gdzie z punktów zrzutowanych z wielu zdjęć z różnych kierunków uzyskano wielokąty oznaczające okna (Rysunek 64).

a)



b)



c) 

Rysunek 64 Przykład działania metodyki ujednolicenia wyników segmentacji okien z wielu kierunków: a) punkty zrzutowane na model 3D z usuniętymi wartościami odstającymi i pogrupowane, b) wielokąty utworzone na podstawie punktów, c) wyświetlone na tle modelu referencyjnego BIM.

Jak widać powyżej prostokąty oznaczające okna uzyskane jako wynik rzutowania segmentów nie są idealnie dopasowane względem modelu referencyjnego (BIM). Jednakże jak wykazano taka metodyka może stanowić uzupełnienie modeli budynków o dodatkową warstwę informacyjną. W zależności od zastosowania i potrzeb oraz tego jak docelowy użytkownik korzystałby z takiej bazy danych możliwe jest przechowywanie zarówno okien w postaci poligonów 3D, jak i informacji o położeniu narożników, czy też współrzędnych środka okna i wysokości oraz szerokości.

Analiza statystyczna wykazała, że odległości między położeniem okna na fasadzie budynku między danymi referencyjnymi BIM są dość duże. Jak widać poniżej (Tabela 25) dla budynku Wydziału Transportu wartość średnia dla wszystkich okien ukształtowała się na poziomie 28cm. Dla Wydziału EITI dokładność połączonych wyników segmentacji podobnie jak w przypadku wyników szczątkowych dla pojedynczych zdjęć wykazała wyższe dokładności. Średni błąd położenia okna na fasadzie wyniósł 16cm. Wartości te były liczone między wierzchołkami uzyskanych poligonów jako wyniku segmentacji a narożnikami warstwy referencyjnej.

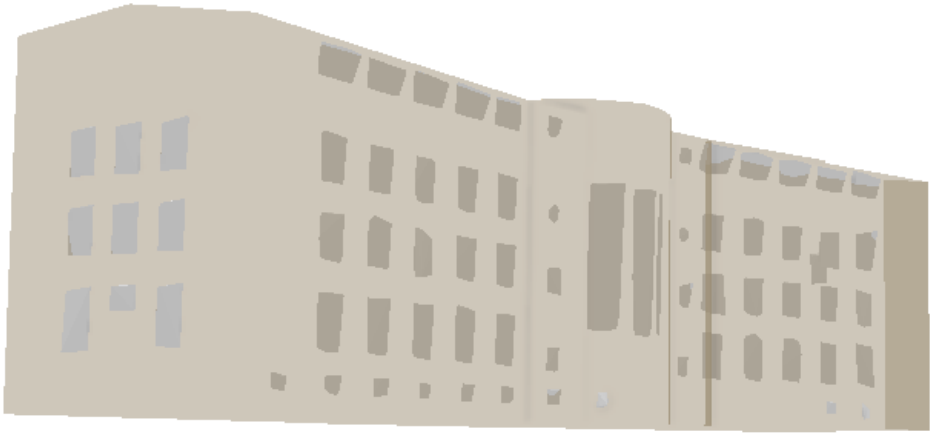
Tabela 25 Wyniki odległości między referencyjnym położeniem okien na fasadach budynków z modeli BIM, a wyznaczonymi na podstawie wyników segmentacji na lotniczych zdjęciach ukośnych i połączenia wyników z wielu kierunków.

	Średnia różnica położenia okna uzyskanego w wyniku segmentacji względem referencji [m]
Budynek Wydziału Transportu	0,28
Budynek Wydziały EITI	0,16

Ponadto warto zauważyć, że mimo iż wcześniejsze wyniki dla budynku Transportu były dużo gorsze, to nadliczbowość obserwacji z wielu kierunków pozwoliła uzyskać poprawne obrysy okien (Rysunek 65). Dzięki temu można ponownie podkreślić zalety jakie płyną w bloku, co pozwala na uzyskanie dobrego wyniku nawet wtedy, gdy na pojedynczych zdjęciach model nie poradził sobie dobrze z segmentacją.

a)

b)



Rysunek 65 Wizualizacja fasady budynku Wydziału Transportu od strony zachodniej: a) model referencyjny BIM, b) model LoD2 pobrany z zasobu geodezyjnego z wyświetlonymi poligonami 3D okien uzyskanych w wyniku łączenia segmentacji z wielu kierunków.

Ostatnim etapem, o który mogłyby zostać poszerzone analizy i badania to poprawa końcowego wyniku, gdzie uwzględniona byłaby średnia wielkość okien na budynku pozyskana na podstawie normalizacji rozmiarów prostokątów (oparta o wzorzec dla okien danego budynku). Implementacja takiej parametryzacji mogłaby jeszcze bardziej uprościć przechowywanie informacji w bazie danych, gdyż mogłyby być przechowywane tylko centroidy okien i informacja o wymiarach.

11.Wnioski

Celem niniejszej pracy było opracowanie metodyki wykrywania obiektów na ukośnych zdjęciach lotniczych z wykorzystaniem metod głębokiego uczenia, a dokładniej konwolucyjnych sieci neuronowych w celu zasilania miejskich baz danych.

Badania przeprowadzone w ramach prac badawczych wykazały przydatność wykorzystania metod głębokiego uczenia do detekcji i segmentacji wybranych obiektów przestrzeni miejskiej tzn. latarni ulicznych i okien na ukośnych zdjęciach lotniczych. Koncentracja na wykrywaniu tych dwóch obiektów była uzasadniona zarówno jeśli chodzi o znaczenie w odniesieniu do przestrzeni miejskiej, jak i ich zróżnicowanie pod względem charakterystyki i metodyki wykrywania oraz lokalizacji. Ponadto latarnie uliczne i okna są reprezentatywne dla szerszego zakresu klas obiektów.

Pierwsza część badań dotyczyła doboru najlepszego źródła danych oraz oceny ich jakości i dokładności pod kątem wykorzystania do wykrywania wybranych obiektów miejskich z wykorzystaniem modeli głębokiego uczenia. W przyjętej metodyce wykorzystano ukośne zdjęcia lotnicze pozyskane sensorem Leica CityMapper2 dla obszarów miejskich. Eksperymenty potwierdziły, że zdjęcia tego typu umożliwiają tworzenie szczegółowej warstwy informacyjnej 3D dla przestrzeni miejskiej z wykorzystaniem modeli takich jak SAM i YOLO.

Szeroko poruszona została tematyka problemów związanych z dostępnością zbiorów uczących do detekcji i segmentacji na ukośnych zdjęciach lotniczych. Wyniki eksperymentów dotyczących analizy kompletności chmur punktów z gęstego dopasowania obrazów jako potencjalnego wsparcia w automatyzacji procesu tworzenia zbiorów uczących wykazały, że wykorzystanie reprezentacji tego samego obiektu z uwzględnieniem różnych typów danych jest podejściem, które może przyspieszyć manualne etykietowanie. W tej części wykorzystano opracowaną wcześniej metodykę, która została w pełni opisana i opublikowana w artykule autorki (Zachar i in., 2022). Metodyka zakłada wykorzystanie fragmentów wyciętych chmur punktów dla obiektów zainteresowania jako danych wejściowych do tworzenia zbioru treningowego do detekcji na zdjęciach ukośnych. Analiza ta pozwoliła zauważyć problemy w pełnym odwzorowaniu latarni ulicznych w chmurach punktów z gęstego dopasowania obrazów. Wykazane zostało także, że dane ALS pozyskane synchronicznie ze zdjęciami są dobrym źródłem do utworzenia kompletnego zestawu referencyjnego do wykrywania na ukośnych zdjęciach lotniczych. W efekcie utworzony zbiór składa się ze zdjęć i informacji o ramkach ograniczających obiekty.

Kolejna część dotyczyła analizy wykonalności zadania detekcji latarni z wykorzystaniem sieci YOLO wytrenowanej na podstawie zbioru, który był efektem wcześniejszej części, wygenerowanym w zautomatyzowany sposób na podstawie wyciętych fragmentów chmur punktów. Wyniki zarówno na zbiorze walidacyjnym, jak i testowym były dobre, co wskazuje na spójną wydajność. Wartości dokładności i precyzji wyniosły ok. 98%, co świadczy o dużej skuteczności. Potwierdziło to zatem założenia, że metody głębokiego uczenia wykazują wysokie dokładności w zadaniu detekcji smukłych, wysokich obiektów takich jak latarnie uliczne, a zdjęcia ukośne są dobrym źródłem danych do rozpoznawania obiektów o takiej charakterystyce.

Oprócz detekcji z wykorzystaniem modelu YOLO przeprowadzono badania dotyczące segmentacji latarni modelem SAM. Uzyskane wyniki pokazały, że lepszą metodą do identyfikacji tego typu obiektów są detektory, gdyż skupiają się na rozpoznaniu i lokalizacji całego obiektu. Segmentacja zaś, która opiera się na dokładnym przypisaniu każdego piksela do danego obiektu okazała się mniej efektywnym podejściem w tym przypadku. Tak więc w zastosowaniach, gdzie precyzyjne zdefiniowanie konturów nie jest wymagane (dla latarni ulicznych), detekcja zapewnia lepsze dokładności niż segmentacja.

Drugim analizowanym obiektem miejskim były okna. W tym przypadku podejściem, które zostało wybrane do rozpoznawania okien była segmentacja obrazów z wykorzystaniem modelu SAM na dwóch obszarach testowych. Dla danych z Bordeaux model uzyskał dokładności na poziomie 78% i segmentował obiekty z precyzją 74%. Natomiast w przypadku Warszawy wyniki dla budynku wydziału EITI ukształtowały się na poziomie ok. 92%. Znacznie gorzej SAM poradził sobie z segmentacją wydziału Transportu, osiągając dokładność ok. 63%. Jednakże biorąc pod uwagę fakt, iż model SAM, który nie był dodatkowo douczany, a bazowano na udostępnionych wagach można uznać, że osiągnięte wyniki świadczą o dużym potencjale takiego podejścia. Ponadto w zastosowaniach, gdzie precyzyjne zdefiniowanie konturów nie jest wymagane (dla latarni ulicznych), detekcja zapewnia lepszy stosunek dokładności niż segmentacja.

Rozbudowane eksperymenty potwierdziły założenia, że modele uczenia głębokiego takie jak YOLO i SAM radzą sobie w zadaniach detekcji i segmentacji, osiągając wysokie dokładności. Podejście, w którym bazuje się jedynie na użyciu modelu przetrenowanego i przeniesieniu go do nowej domeny może działać dobrze. W przypadku sieci, które zostały zaprojektowane tak, by były jak najbardziej uniwersalne (np. SAM) i była możliwa ich generalizacja do nowego zadania możliwa jest adaptacja bez konieczności modyfikacji jego wag.

Natomiast wykorzystanie metod transfer learning i data augmentation jest zasadne, gdy konieczne jest uzyskanie wyników o dużo wyższej dokładności, czego potwierdzeniem są wyniki detekcji latarni z wykorzystaniem modelu YOLO, gdzie obie metody były zaadaptowane, a skuteczność wykrywania była dużo wyższa w porównaniu z innymi wariantami bez trenowania.

Można więc podsumować, że rozwiązanie gdzie nie doucza się modelu sprawdzi się w takim celu jak przyspieszenie i zautomatyzowanie tworzenia zbioru treningowego, natomiast kiedy model przeznaczony jest pod specyficzne zastosowanie konieczne jest dotrenowanie i dostosowanie wag pod konkretne przeznaczenie.

Ostatnia przeprowadzona analiza dotycząca metodyki wyznaczania lokalizacji potwierdziła, że możliwe jest określenie położenia obiektów w terenowym układzie współrzędnych na podstawie wyników detekcji i segmentacji z modeli DL, a wyniki mogą stanowić dodatkową warstwę informacyjną w bazie danych o obiektach miejskich. Dla latarni ulicznych zweryfikowano dwie metody: fotogrametryczne wcięcie w przód oraz transformację współrzędnych punktu z układu współrzędnych pikselowych zdjęcia do terenowego układu współrzędnych na podstawie warunku kolinearności z wykorzystaniem NMT. Dla metody rzutowania średnia wartość odległości między referencyjną lokalizacją latarni a wynikową z detekcji na zdjęciach lotniczych wyniosła 14cm, zaś dla metody realizującej wcięcie – 13cm. Dla okien wykorzystano jedno podejście związane z rzutowaniem na model mesh 3D. Uzyskano wyniki na poziomie 16cm dla budynku Wydziału EITI i 28cm dla budynku Wydziału Transportu jeśli chodzi o błąd położenia okna na fasadzie względem referencyjnego położenia z modeli BIM.

Wykazano zatem praktyczne zastosowanie wyników detekcji latarni i segmentacji okien tzn. automatyzację zasilania baz danych obiektów miejskich, realizując geolokalizację rozpoznanych obiektów.

Pomimo, iż nie wykazano dokładności, które byłyby zadowalające w warunkach produkcyjnych, to wyniki z poszczególnych etapów pokazują, że zaproponowany ciąg technologiczny może przede wszystkim zautomatyzować pracę i odpowiedzieć na postawione wyzwania badawcze. Należy szczególnie zwrócić uwagę na wykorzystywanie modeli z wytrenowanymi wagami bez dodatkowego uczenia. Ponadto wartości mówiące o czułości algorytmu mogą stanowić podstawę do stwierdzenia, że model popełnił małą liczbę błędów podczas detekcji czy segmentacji. Dodatkowe dotrenowanie modelu nawet na małej próbce

danych prawdopodobnie znacząco polepszyłoby statystyki dokładnościowe. Jednak biorąc pod uwagę fakt, iż głównym celem badań nie było wytrenowanie czy opracowanie jak najbardziej dokładnego modelu do wykrywania latarni ulicznych i segmentacji okien na zdjęciach, a analiza potencjału wykorzystania metod głębokiego uczenia do zadań pozwalających zautomatyzować i przyspieszyć prace związane z uzupełnianiem baz danych obiektów miejskich to można uznać, że wykonane eksperymenty odpowiadają na poszczególne postawione w początkowej fazie badań pytania.

Podsumowując szereg wykonanych eksperymentów pozwolił na osiągnięcie celów szczegółowych pracy, a także uzyskanie odpowiedzi na postawione pytania badawcze. Wyniki pracy dowodzą udowodnienia postawionej na wstępie tezy.

Bibliografia:

Abdelfattah, R., Wang, X., & Wang, S. (2020). Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines. In *Proceedings of the Asian Conference on Computer Vision*.

Adam, J. M., Liu, W., Zang, Y., Afzal, M. K., Bello, S. A., Muhammad, A. U., ... & Li, J. (2023). Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey. *International Journal of Applied Earth Observation and Geoinformation*, *121*, 103365.

Ahmed, I., Ahmad, M., Chehri, A., Hassan, M. M., & Jeon, G. (2022). IoT enabled deep learning based framework for multiple object detection in remote sensing images. *Remote Sensing*, *14*(16), 4107.

Aijazi, A. K., Checchin, P., & Trassoudaine, L. (2013). Segmentation based classification of 3D urban point clouds: A super-voxel based approach with evaluation. *Remote Sensing*, *5*(4), 1624-1650.

Amirebrahimi, S., Rajabifard, A., Mendis, P., & Ngo, T. (2016). A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration. *International Journal of Digital Earth*, *9*(4), 363-386.

Amiri, N., Heurich, M., Krzystek, P., & Skidmore, A. K. (2018). Feature relevance assessment of multispectral airborne LiDAR data for tree species classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*(3), 31-34.

Axelsson, A., Lindberg, E., & Olsson, H. (2018). Exploring multispectral ALS data for tree species classification. *Remote Sensing*, *10*(2), 183.

Bacher, U. (2022). Hybrid aerial sensor data as basis for a geospatial digital twin. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 653-659.

Benedek, C., Descombes, X., & Zerubia, J. (2011). Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(1), 33-50.

Bozcan, I., & Kayacan, E. (2020, May). Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8504-8510). IEEE.

Braun, A., & Borrmann, A. (2019). Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning. *Automation in Construction*, *106*, 102879.

Cao, M. T. (2023). Drone-assisted segmentation of tile peeling on building façades using a deep learning model. *Journal of Building Engineering*, *80*, 108063.

Carnegie, A. J., Eslick, H., Barber, P., Nagel, M., & Stone, C. (2023). Airborne multispectral imagery and deep learning for biosecurity surveillance of invasive forest pests in urban landscapes. *Urban Forestry & Urban Greening*, *81*, 127859.

Chen, B., & Miao, X. (2020). Distribution line pole detection and counting based on YOLO using UAV inspection line video. *Journal of Electrical Engineering & Technology*, *15*(1), 441-448.

Chen, K., Wu, M., Liu, J., & Zhang, C. (2020). FGSD: A dataset for fine-grained ship detection in high resolution satellite images. *arXiv preprint arXiv:2003.06832*.

Chen, L., Chang, J., Xu, J., & Yang, Z. (2023). Automatic measurement of inclination angle of utility poles using 2D image and 3D point cloud. *Applied Sciences*, *13*(3), 1688.

Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., ... & Soibelman, L. (2022). Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*.

Cheng, G., Han, J., Zhou, P., & Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, *98*, 119-132.

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., & Han, J. (2023). Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.

Deokuliar, H. (2023). Object detection using drone-based aerial imagery in urban planning and land development projects. A Thesis in Data Analytics.

Ding, J., Xue, N., Xia, G. S., Bai, X., Yang, W., Yang, M. Y., ... & Zhang, L. (2021). Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, *44*(11), 7778-7796.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Farajzadeh, Z., Saadatseresht, M., & Alidoost, F. (2023). Automatic building extraction from UAV-based images and DSMs using deep learning. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, *10*, 171-177.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, *61*(3), 399-409.

Gajski, D., & Dzięgielewska-Gajski, K. (2023). Nadir & Oblique Aerial Imagery–new possibilities in 3D mapping. *GIS Odyssey Journal*, *3*(1), 137-147.

Gallagher, J. E., Gogia, A., & Oughton, E. J. (2024). A Multispectral Automated Transfer Technique (MATT) for machine-driven image labeling utilizing the Segment Anything Model (SAM). *arXiv preprint arXiv:2402.11413*.

Gao, W., Nan, L., Boom, B., & Ledoux, H. (2021). SUM: A benchmark dataset of semantic urban meshes. *ISPRS Journal of Photogrammetry and Remote Sensing*, *179*, 108-120.

Garg, P., Chakravarthy, A. S., Mandal, M., Narang, P., Chamola, V., & Guizani, M. (2021). Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. *ACM Transactions on Internet Technology (TOIT)*, *21*(3), 1-18.

Gerke, M., Nex, F., Remondino, F., Jacobsen, K., Kremer, J., Karel, W., ... & Ostrowski, W. (2016). Orientation of oblique airborne image sets-experiences from the ISPRS/EUROSDR benchmark on multi-platform photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 41-B1*, *41*, 185-191.

Gomes, M., Silva, J., Gonçalves, D., Zamboni, P., Perez, J., Batista, E., ... & Gonçalves, W. (2020). Mapping utility poles in aerial orthoimages using atss deep learning method. *Sensors*, *20*(21), 6070.

Griffiths, D., & Boehm, J. (2019). SynthCity: A large scale synthetic point cloud. *arXiv preprint arXiv:1907.04758*.

Grzeczkowicz, G., & Vallet, B. (2023). Semantic segmentation of urban textured meshes through point sampling. *arXiv preprint arXiv:2302.10635*.

Guan, H., Lei, X., Yu, Y., Zhao, H., Peng, D., Junior, J. M., & Li, J. (2022). Road marking extraction in UAV imagery using attentive capsule feature pyramid network. *International Journal of Applied Earth Observation and Geoinformation*, *107*, 102677.

Haala, N., Rothermel, M., & Cavegn, S. (2015, March). Extracting 3D urban models from oblique aerial images. In *2015 Joint Urban Remote Sensing Event (JURSE)* (pp. 1-4). IEEE.

Han, Y., Yang, X., Pu, T., & Peng, Z. (2021). Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-18.

Haroon, M., Shahzad, M., & Fraz, M. M. (2020). Multisized object detection using spaceborne optical imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3032-3046.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).

Heitz, G., & Koller, D. (2008). Learning spatial context: Using stuff to find things. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10* (pp. 30-43). Springer Berlin Heidelberg.

Heo, W. Y., Kim, S., Yoon, D., Jeong, J., & Sung, H. (2020, April). Deep learning based moving object detection for oblique images without future frames. In *Automatic Target Recognition XXX* (Vol. 11394, p. 1139403). SPIE.

Hoeser, T., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing*, *12*(10), 1667.

Hoeser, T., Bachofer, F., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. *Remote Sensing*, *12*(18), 3053.

Hsieh, M. R., Lin, Y. L., & Hsu, W. H. (2017). Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision* (pp. 4145-4153).

Huang, H., Savkin, A. V., & Huang, C. (2021). Decentralized autonomous navigation of a UAV network for road traffic monitoring. *IEEE Transactions on Aerospace and Electronic Systems*, *57*(4), 2558-2564.

Huang, J., & You, S. (2015, May). Pole-like object detection and classification from urban point clouds. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3032-3038). IEEE.

Huang, S. (2019). *Building segmentation in oblique aerial imagery* (Master's thesis, University of Twente).

Huang, S., Nex, F., Lin, Y., & Yang, M. Y. (2019). Semantic segmentation of building in airborne images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 35-42.

Jin, J., Gubbi, J., Marusic, S., & Palaniswami, M. (2014). An information framework for creating a smart city through internet of things. *IEEE Internet of Things journal*, *1*(2), 112-121.

Kang, Z., Yang, J., Zhong, R., Wu, Y., Shi, Z., & Lindenbergh, R. (2018). Voxel-based extraction and classification of 3-D pole-like objects from mobile LiDAR point cloud data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(11), 4287-4298.

Khatua, A., Bhattacharya, A., Goswami, A. K., & Aithal, B. H. (2024). Developing approaches in building classification and extraction with synergy of YOLOV8 and SAM models. *Spatial Information Research*, 1-20.

Kim, S., Zadeh, P. A., Staub-French, S., Froese, T., & Cavka, B. T. (2016). Assessment of the impact of window size, position and orientation on building energy load using BIM. *Procedia Engineering*, *145*, 1424-1431.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015-4026).

Kisantal, M. (2019). Augmentation for Small Object Detection. *arXiv preprint arXiv:1902.07296*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., ... & McCord, B. (2018). xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*.

Laupheimer, D., & Haala, N. (2021). Juggling with representations: On the information transfer between imagery, point clouds, and meshes for multi-modal semantics. *ISPRS Journal of Photogrammetry and Remote Sensing*, *176*, 55-68.

Lee, C., Soedarmadji, S., Anderson, M., Clark, A. J., & Chung, S. J. (2024). Semantics from Space: Satellite-Guided Thermal Semantic Segmentation Annotation for Aerial Field Robots. *arXiv preprint arXiv:2403.14056*.

Lee, G., Hwang, J., & Cho, S. (2021). A novel index to detect vegetation in urban areas using UAV-based multispectral images. *Applied sciences*, *11*(8), 3472.

Li, D. (2013). *Optimising detection of road furniture (pole-like objects) in mobile laser scanning data* (Master's thesis, University of Twente).

Li, J., Qu, C., & Shao, J. (2017, November). Ship detection in SAR images based on an improved faster R-CNN. In *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)* (pp. 1-6). IEEE.

Li, K., Wan, G., Cheng, G., Meng, L., & Han, J. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, *159*, 296-307.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, *11*(4), 403.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935-2947.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y., & Hoi, S. C. (2020). DeepFacade: A deep learning approach to facade parsing with symmetric loss. *IEEE Transactions on Multimedia*, 22(12), 3153-3165.

Liu, H., Zhang, J., Zhu, J., & Hoi, S. C. (2017). Deepfacade: A deep learning approach to facade parsing. IJCAI.

Liu, K., & Mattyus, G. (2015). Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, *12*(9), 1938-1942.

Liu, Z., Yuan, L., Weng, L., & Yang, Y. (2017, February). A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods* (Vol. 2, pp. 324-331). SciTePress.

Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cué La Rosa, L., Marcato Junior, J., Martins, J., ... & Liesenberg, V. (2020). Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery. *Sensors*, *20*(2), 563.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., & Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277-293.

Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M. O., & Huang, B. (2024). Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geoscience and Remote Sensing*.

Mao, Z., Huang, X., Niu, W., Wang, X., Hou, Z., & Zhang, F. (2023). Improved instance segmentation for slender urban road facility extraction using oblique aerial images. *International Journal of Applied Earth Observation and Geoinformation*, *121*, 103362.

Mao, Z., Huang, X., Xiang, H., Gong, Y., Zhang, F., & Tang, J. (2023). Glass façade segmentation and repair for aerial photogrammetric 3D building models with multiple constraints. *International Journal of Applied Earth Observation and Geoinformation*, *118*, 103242.

Mao, Z., Zhang, F., Huang, X., Jia, X., Gong, Y., & Zou, Q. (2021). *Deep Neural Networks for Road Sign Detection and Embedded Modeling Using Oblique Aerial Images. Remote Sens. 2021, 13, 879*.

Mathias, M., Martinović, A., & Van Gool, L. (2016). ATLAS: A three-layered approach to facade parsing. *International Journal of Computer Vision*, *118*, 22-48.

Menard, S.2018. Applied Logistic Regression Analysis. New York: SAGE

Michałowska, M., & Rapiński, J. (2021). A review of tree species classification based on airborne LiDAR data and applied classifiers. *Remote Sensing*, *13*(3), 353.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(7), 3523-3542.

Moudrý, V., Cord, A. F., Gábor, L., Laurin, G. V., Barták, V., Gdulová, K., ... & Wild, J. (2023). Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. *Diversity and Distributions*, *29*(1), 39-50.

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for UAV tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 445-461). Springer International Publishing.

Muhmad Kamarulzaman, A. M., Wan Mohd Jaafar, W. S., Mohd Said, M. N., Saad, S. N. M., & Mohan, M. (2023). UAV Implementations in Urban Planning and Related Sectors of Rapidly Developing Nations: A Review and Future Perspectives for Malaysia. *Remote Sensing*, *15*(11), 2845.

Mundhenk, T. N., Konjevod, G., Sakla, W. A., & Boakye, K. (2016). A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (pp. 785-800). Springer International Publishing.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote sensing of environment*, *80*(1), 88-99.

Prado Osco, L., Wu, Q., Lopes de Lemos, E., Nunes Gonçalves, W., Marques Ramos, A. P., Li, J., & Marcato Junior, J. (2023). The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot. *arXiv e-prints*, arXiv-2306.

Pu, S., Rutzinger, M., Vosselman, G., & Elberink, S. O. (2011). Recognizing basic structures from mobile laser scanning data for road inventory studies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(6), S28-S39.

Qin, R., & Gruen, A. (2021). The role of machine intelligence in photogrammetric 3D modelling – an overview and perspectives. *International Journal of Digital Earth*, *14*(1), 15-31.

Razakarivony, S., & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, *34*, 187-203.

Redmon, J. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Rizzoli, G., Barbato, F., Caligiuri, M., & Zanuttigh, P. (2023). Syndrone-multi-modal uav dataset for urban scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2210-2220).

Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14* (pp. 549-565). Springer International Publishing.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3234-3243).

Schlosser, A. D., Szabó, G., Bertalan, L., Varga, Z., Enyedi, P., & Szabó, S. (2020). Building extraction using orthophotos and dense point cloud derived from visual band aerial imagery based on machine learning and segmentation. *Remote Sensing*, *12*(15), 2397.

Schmitz, M., & Mayer, H. (2016). A convolutional network for semantic facade segmentation and interpretation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *41*, 709-715.

Shermeyer, J., Hossler, T., Van Etten, A., Hogan, D., Lewis, R., & Kim, D. (2021). Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 207-217).

Shi, Z., Kang, Z., Lin, Y., Liu, Y., & Chen, W. (2018). Automatic recognition of pole-like objects from mobile laser scanning point clouds. *Remote Sensing*, *10*(12), 1891.

Song, J., Gao, S., Zhu, Y., & Ma, C. (2019). A survey of remote sensing image classification based on CNNs. *Big earth data*, *3*(3), 232-254.

Sultan, R. I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., & Zhu, D. (2023). GeoSAM: Fine-tuning SAM with sparse and dense visual prompting for automated segmentation of mobility infrastructure. *arXiv preprint arXiv:2311.11319*.

Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., ... & Fu, K. (2022). FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *184*, 116-130.

Tahir, N. U. A., Long, Z., Zhang, Z., Asim, M., & ELAffendi, M. (2024). PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8. *Drones*, *8*(3), 84.

Taye, M. M. (2023). Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. *Computation*, *11*(3), 52.

Toschi, I., Ramos, M. M., Nocerino, E., Menna, F., Remondino, F., Moe, K., ... & Fassi, F. (2017). Oblique photogrammetry supporting 3D urban reconstruction of complex scenarios. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 519-526.

Toschi, I., Remondino, F., Rothe, R., & Klimek, K. (2018). Combining airborne oblique camera and LiDAR sensors: Investigation and new perspectives. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 437-444.

Toschi, I., Remondino, F., Hauck, T., & Wenzel, K. (2019). When photogrammetry meets LiDAR: towards the airborne hybrid era. *GIM INTERNATIONAL*, 17-21.

Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). Ieee.

Wang, G., Chen, Y., An, P., Hong, H., Hu, J., & Huang, T. (2023). UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. *Sensors*, *23*(16), 7190.

Wang, L. (Ed.). (2005). *Support vector machines: theory and applications* (Vol. 177). Springer Science & Business Media.

Wang, L., Hu, H., Shang, Q., Xu, B., & Zhu, Q. (2023). StructuredMesh: 3D Structured Optimization of Fa\c {c} ade Components on Photogrammetric Mesh Models using Binary Integer Programming. *arXiv preprint arXiv:2306.04184*.

Wang, R., Huang, S., & Yang, H. (2023). Building3d: A urban-scale dataset and benchmarks for learning roof structures from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 20076-20086).

Wang, Z., Yang, L., Sheng, Y., & Shen, M. (2021). *Pole-Like Objects Segmentation and Multiscale Classification-Based Fusion from Mobile Point Clouds in Road Scenes. Remote Sens. 2021, 13, 4382*.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., ... & Bai, X. (2019). isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 28-37).

Weinmann, M., Weinmann, M., Mallet, C., & Brédif, M. (2017). A classification-segmentation framework for the detection of individual trees in dense MMS point cloud data acquired in urban areas. *Remote Sensing*, *9*(3), 277.

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A. V., McPherson, S., Shermeyer, J., ... & Tang, H. (2019). Spacenet mvoi: A multi-view overhead imagery dataset. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 992-1001).

White, J. C., Stepper, C., Tompalski, P., Coops, N. C., & Wulder, M. A. (2015). Comparing ALS and image-based point cloud metrics and modelled forest inventory attributes in a complex coastal forest environment. *Forests*, *6*(10), 3704-3732.

Wilk, Ł., Mielczarek, D., Ostrowski, W., Dominik, W., & Krawczyk, J. (2022). Semantic urban mesh segmentation based on aerial oblique images and point clouds using deep learning. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 485-491.

Wu, T., & Dong, Y. (2023). YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. *Applied Sciences*, *13*(24), 12977.

Xian, S. U. N., Zhirui, W. A. N. G., Yuanrui, S. U. N., Wenhui, D. I. A. O., Yue, Z. H. A. N. G., & Kun, F. U. (2019). AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *雷达学报*, *8*(6), 852-863.

Xiao, Z., Liu, Q., Tang, G., & Zhai, X. (2015). Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *International Journal of Remote Sensing*, *36*(2), 618-644.

Xie, J., Kiefel, M., Sun, M. T., & Geiger, A. (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 3688-3697).

Xiong, Z., Zhang, F., Wang, Y., Shi, Y., & Zhu, X. X. (2022). Earthnets: Empowering ai in earth observation. *arXiv preprint arXiv:2210.04936*.

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., & Xia, G. S. (2022). Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 79-93.

Yan, W. Y., Morsy, S., Shaker, A., & Tulloch, M. (2016). Automatic extraction of highway light poles and towers from mobile LiDAR data. *Optics & Laser Technology*, *77*, 162-168.

Yang, C., Zhang, F., Gao, Y., Mao, Z., Li, L., & Huang, X. (2021). Moving car recognition and removal for 3D urban modelling using oblique images. *Remote Sensing*, *13*(17), 3458.

Yang, M. Y., Liao, W., Li, X., & Rosenhahn, B. (2018, October). Deep learning for vehicle detection in aerial images. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3079-3083). IEEE.

Yang, Z., Liu, Y., Liu, L., Tang, X., Xie, J., & Gao, X. (2019). Detecting small objects in urban settings using SlimNet model. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(11), 8445-8457.

Yi, H., Liu, B., Zhao, B., & Liu, E. (2023). Small object detection algorithm based on improved YOLOv8 for remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Yu, H., Li, G., Zhang, W., Huang, Q., Du, D., Tian, Q., & Sebe, N. (2020). The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, *128*, 1141-1159.

Yu, W. Q.， Cheng G.， Wang M. J.， Yao Y. Q.， Xie X. X.， Yao X. X . and Han J. W. 2023. MAR20 ： A benchmark for military aircraft recognition in remote sensing images. National Remote Sensing Bulletin, 27 (12): 2688-2696 DOI: 10.11834/jrs.20222139.

Yu, X., Hyyppä, J., Litkey, P., Kaartinen, H., Vastaranta, M., & Holopainen, M. (2017). Single-sensor solution to tree species classification using multispectral airborne laser scanning. *Remote Sensing*, *9*(2), 108.

Zachar, P., Bakuła, K., & Ostrowski, W. (2023). Cenagis-Als Benchmark-New Proposal for Dense ALS Benchmark Based on the Review of Datasets and Benchmarks for 3d Point Cloud Segmentation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48, 227-234.

Zachar, P., Kurczyński, Z., & Ostrowski, W. (2022). Application of machine learning for object detection in oblique aerial images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 657-663.

Zachar, P., Ostrowski, W., Płatek-Żak, A., & Kurczyński, Z. (2022). The Influence of Point Cloud Accuracy from Image Matching on Automatic Preparation of Training Datasets for Object Detection in UAV Images. ISPRS International Journal of Geo-Information, 11(11), 565.

Zhang, G., & Zhang, R. (2023). MeshNet-SP: A Semantic Urban 3D Mesh Segmentation Network with Sparse Prior. *Remote Sensing*, *15*(22), 5324.

Zhang, H., Sun, M., Li, Q., Liu, L., Liu, M., & Ji, Y. (2021). An empirical study of multi-scale object detection in high resolution UAV images. *Neurocomputing*, *421*, 173-182.

Zhang, W., Liu, C., Chang, F., & Song, Y. (2020). Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images. *Remote Sensing*, *12*(11), 1760.

Zhang, W., Witharana, C., Li, W., Zhang, C., Li, X., & Parent, J. (2018). Using deep learning to identify utility poles with crossarms and estimate their locations from google street view images. *Sensors*, *18*(8), 2484.

Zhang, X., & Aliaga, D. (2022). RFCNet: Enhancing urban segmentation using regularization, fusion, and completion. *Computer Vision and Image Understanding*, 220, 103435.

Zhang, Y., Yuan, Y., Feng, Y., & Lu, X. (2019). Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(8), 5535-5548.

Zhao, J., Wang, Y., Cao, Y., Guo, M., Huang, X., Zhang, R., ... & Wang, J. (2021). The fusion strategy of 2d and 3d information based on deep learning: A review. *Remote Sensing*, *13*(20), 4029.

Zhao, W., Persello, C., & Stein, A. (2021). Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS journal of photogrammetry and remote sensing*, *175*, 119-131.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015, September). Orientation robust object detection in aerial images using deep convolutional neural network. In 2015 IEEE international conference on image processing (ICIP) (pp. 3735-3739). IEEE.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7380-7399.

Zhuo, X., Fraundorfer, F., Kurz, F., & Reinartz, P. (2018, July). Building detection and segmentation using a CNN with automatically generated training data. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 3461-3464). IEEE.

Zhuo, X., Mönks, M., Esch, T., & Reinartz, P. (2019, May). Facade segmentation from oblique UAV imagery. In *2019 Joint Urban Remote Sensing Event (JURSE)* (pp. 1-4). IEEE.

Zolanvari, S. M., Ruano, S., Rana, A., Cummins, A., Da Silva, R. E., Rahbar, M., & Smolic, A. (2019). DublinCity: Annotated LiDAR point cloud and its applications. *arXiv preprint arXiv:1909.03613*.

Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T. Y., Shlens, J., & Le, Q. V. (2020). Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16* (pp. 566-583). Springer International Publishing.

Zou, Z., & Shi, Z. (2017). Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, *27*(3), 1100-1111.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, *111*(3), 257-276.

Spis rysunków

Spis tabel