ZASTOSOWANIE GŁĘBOKICH SIECI NEURONOWYCH DO DETEKCJI BUDYNKÓW NA ZDJĘCIACH LOTNICZYCH

Mateusz Gomulski

DEFINICJA PROBLEMU BADAWCZEGO

Świadomość dokładnego położenia zabudowań na mapach miast jest niezwykle istotna w wielu dziedzinach życia społeczno-ekonomicznego, do najważniejszych z nich należą:

- urbanistyka,
- socjologia,
- bezpieczeństwo publiczne,
- ubezpieczenia majątkowe,
- reagowanie kryzysowe w obliczu klęsk żywiołowych.

Obecnie stosowane metody identyfikacji położenia budynków na mapach są mało efektywne i bardzo czasochłonne, stąd powstała potrzeba do usprawnienia i automatyzacji tego procesu, a jednym ze sposobów osiągnięcia tego celu jest zastosowanie **głębokich sieci neuronowych** (GSN).

GŁĘBOKIE SIECI NEURONOWE A DETEKCJA BUDYNKÓW

Zadanie detekcji budynków na zdjęciach lotniczych sprowadza się de facto do wskazania dla każdego piksela na danym zdjęciu czy reprezentuje on budynek czy nie.

Głęboka sieć neuronowa ma więc za zadanie wskazanie dla każdego piksela ze zdjęcia wejściowego prawdopodobieństwa, że piksel ten reprezentuje budynek. Następnie pikselom przekraczających ustalony arbitralnie próg prawdopodobieństwa (zwykle 50%) przypisuje się wartość 1, a pozostałym wartość 0.

Taką klasę zadań w nurcie głębokiego uczenia nazywamy **semantyczną segmentacją.**

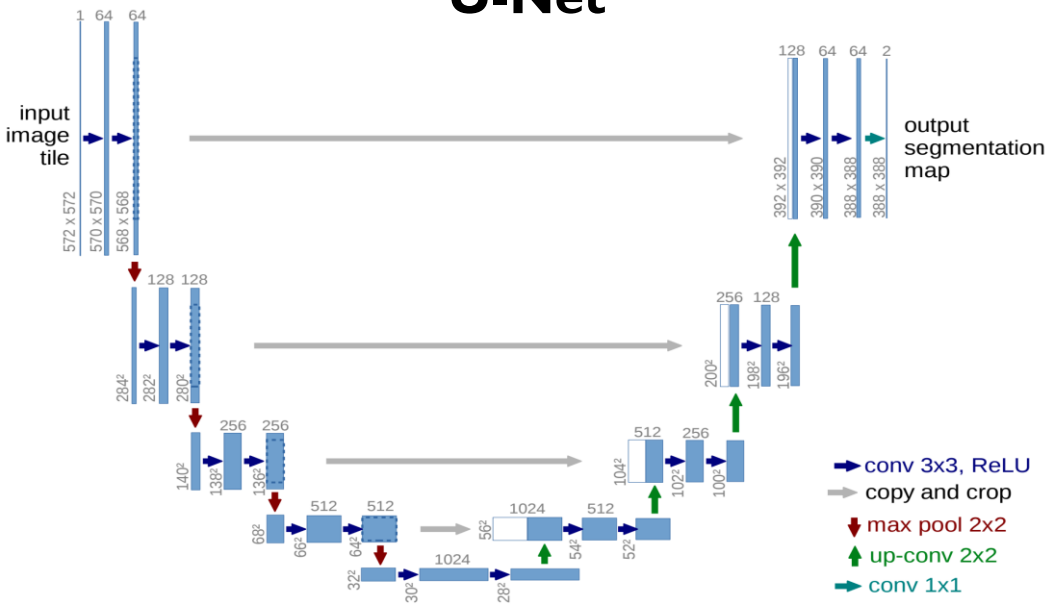
ARCHITEKTURY GSN STOSOWANE DO DETEKCJI BUDYNKÓW

Większość głębokich sieci neuronowych używanych w zadaniach dotyczących detekcji budynków na zdjęciach lotniczych stosuje architektury składające się z dwóch części: enkodera i dekodera.

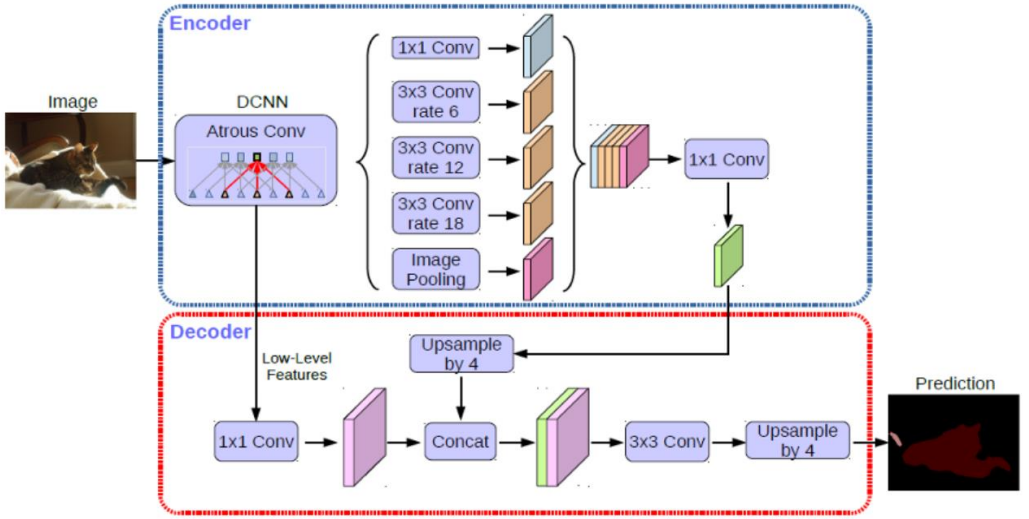
Zadaniem enkodera jest ekstrakcja cech z wejściowego obrazu, natomiast zadaniem dekodera jest propagacja uzyskanej informacji kontekstowej do warstw o wyższej rozdzielczości przestrzennej.

Do najczęściej stosowanych architektur GSN przy detekcji budynków na zdjęciach lotniczych należą: U-Net, DeepLab v3 oraz LinkNet.

U-Net



DeepLab v3



CEL, PRZEDMIOT I METODYKA BADAŃ

Celem omawianego w bieżącej pracy badania własnego było skonstruowanie głębokiej sieci neuronowej, która umożliwi efektywną detekcję zabudowań na zdjęciach lotniczych o wysokiej rozdzielczości.

W tym celu stworzono konwolucyjną sieć neuronową o nazwie ***GML-Net***, która została wytrenowana na zbiorze danych *Inria Aerial Image Labeling Dataset (IAILD)*.

Aby móc określić jakość generowanych przez sieć *GML-Net* predykcji, zdecydowano się na następujące metryki:

- ogólna dokładność (dalej określana angielskim skrótem *OA*),
- wynik F1 (dalej określany angielskim skrótem *F1S*),
- współczynnik podobieństwa Jaccarda (dalej określany angielskim skrótem IoU),
- indeks podobieństwa strukturalnego (dalej określany angielskim skrótem *SSIM*).

DEFINICJA ZBIORÓW DANYCH

Zbiór danych Inria Aerial Image Labeling Dataset (*IAILD*) składa się z 360 zdjęć lotniczych 10 różnych miast świata (po 36 zdjęć każdego z miast). Każde zdjęcie w tym zbiorze ma rozdzielczość 5000x5000 pikseli, trzy kanały przestrzenne i pokrywa teren o powierzchni około 1500x1500 m^2.

Zbiór treningowy został wydzielony ze zbioru *IAILD* – składa się ze 180 zdjęć, po 36 zdjęć z każdego z miast: Austin, Chicago, Hrabstwo Kitsap, Tyrol. Do każdego zdjęcia w zbiorze treningowym przypisana jest maska *ground truth*, która wskazuje które piksele tego zdjęcia reprezentują budynki.

Zbiór walidacyjny został z kolei wydzielony ze zbioru treningowego IAILD jako pięć pierwszych zdjęć (i masek) ww. miast – taka konstrukcja zbioru walidacyjnego *IAILD* jest powszechnie stosowana w literaturze badawczej.

W konsekwencji uzyskano trzy rozłączne zbiory danych:

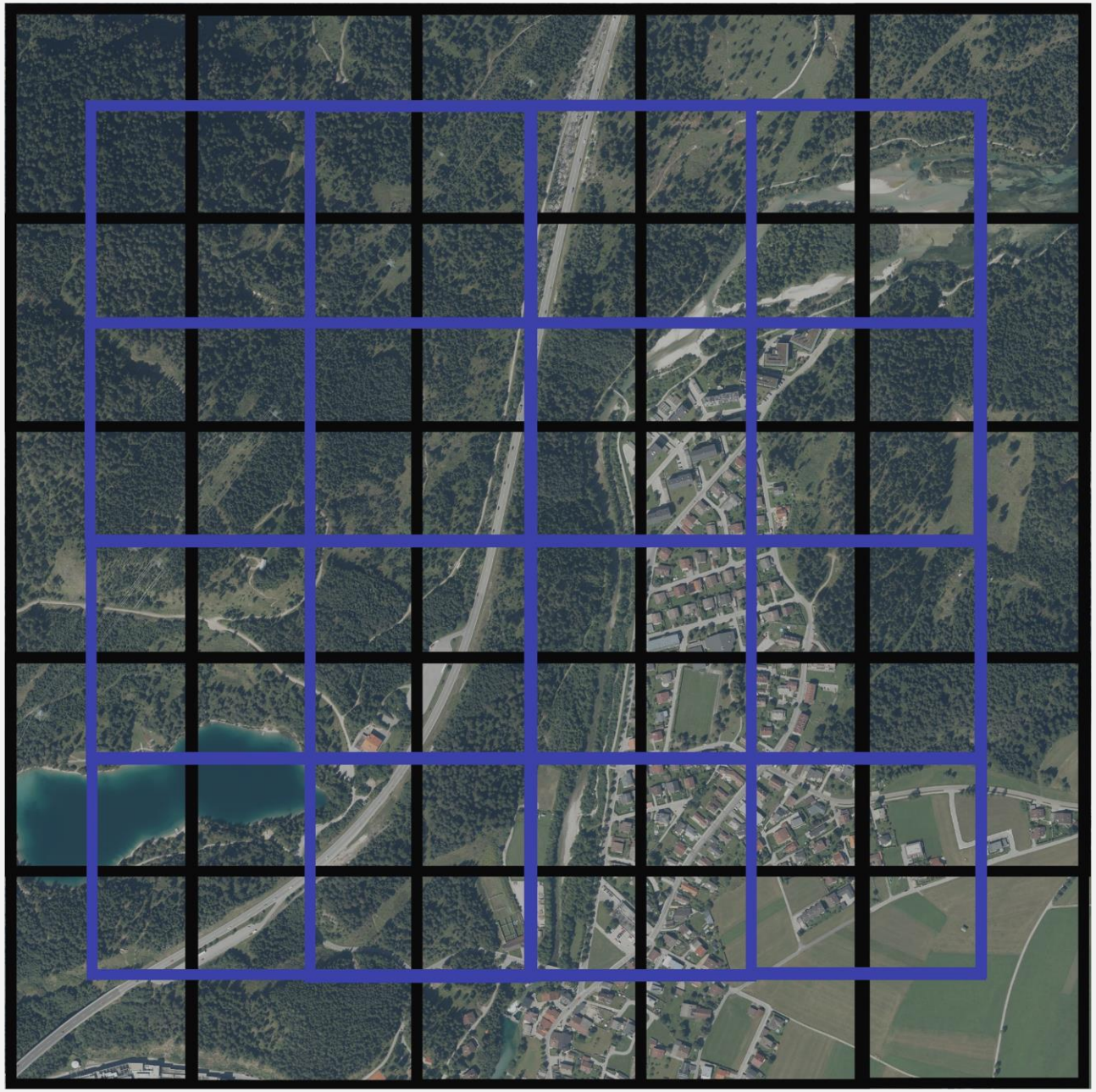
- **zbiór testowy** liczący 180 zdjęć miast: Bellingham, Bloomington, Innsbruck, San Francisco oraz Tyrol Wschodni,
- **zbiór treningowy** liczący 155 zdjęć miast: Austin, Chicago, Hrabstwo Kitsap, Tyrol.
- **zbiór walidacyjny** liczący 25 zdjęć miast: Austin, Chicago, Hrabstwo Kitsap, Tyrol.

DEFINICJA DANYCH UCZĄCYCH

W związku z tym, iż trening głębokiej sieci neuronowej na zdjęciach o rozdzielczości 5000x5000 pikseli nie jest możliwy, zdecydowano się na podział tych zdjęć (oraz ich masek) na 25 równych części, każda o rozdzielczości 1000x1000 pikseli (czarne ramki).

Dodatkowo, żeby nie utracić informacji znajdujących się na krawędziach podzielonych zdjęć (oraz ich masek), zdecydowano się wydzielić kolejnych 16 części znajdujących się na łączeniach każdych czterech sąsiadujących ze sobą bazowych części (niebieskie ramki).

W ten sposób z każdego zdjęcia pobieranych było 41 fragmentów o wymiarach 1000x1000 pikseli, uzyskując tym samym 6355 fragmentów trenujących.



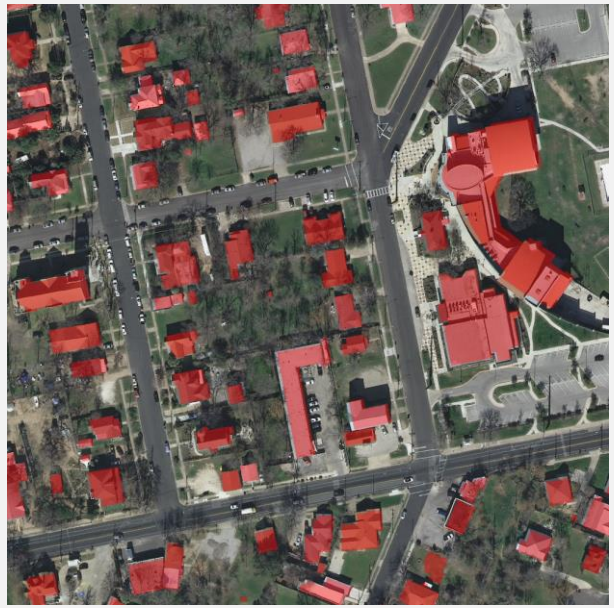
AUGMENTACJE DANYCH UCZĄCYCH

Fragmenty trenujące (1000x1000 pikseli) zostały poddane następującym transformacjom:

- losowy horyzontalny obrót,
- losowy wertykalny obrót,
- rotacja o losowy kąt będący wielokrotnością kąta 30 stopni,
- wybór losowego okna o rozmiarach 256x256 pikseli,
- normalizacja przy użyciu wyliczonych średnich i odchyleń standardowych dla wszystkich zdjęć.

W związku z tym, iż z każdego fragmentu uczącego wybieranych było 18 losowych okien o rozmiarach 256x256 pikseli (rozmiar partii wynosi 18), finalnie uzyskano 114 390 przykładów uczących.

Oryginalne zdjęcie 5000 x 5000



Przykład uczący 256x256

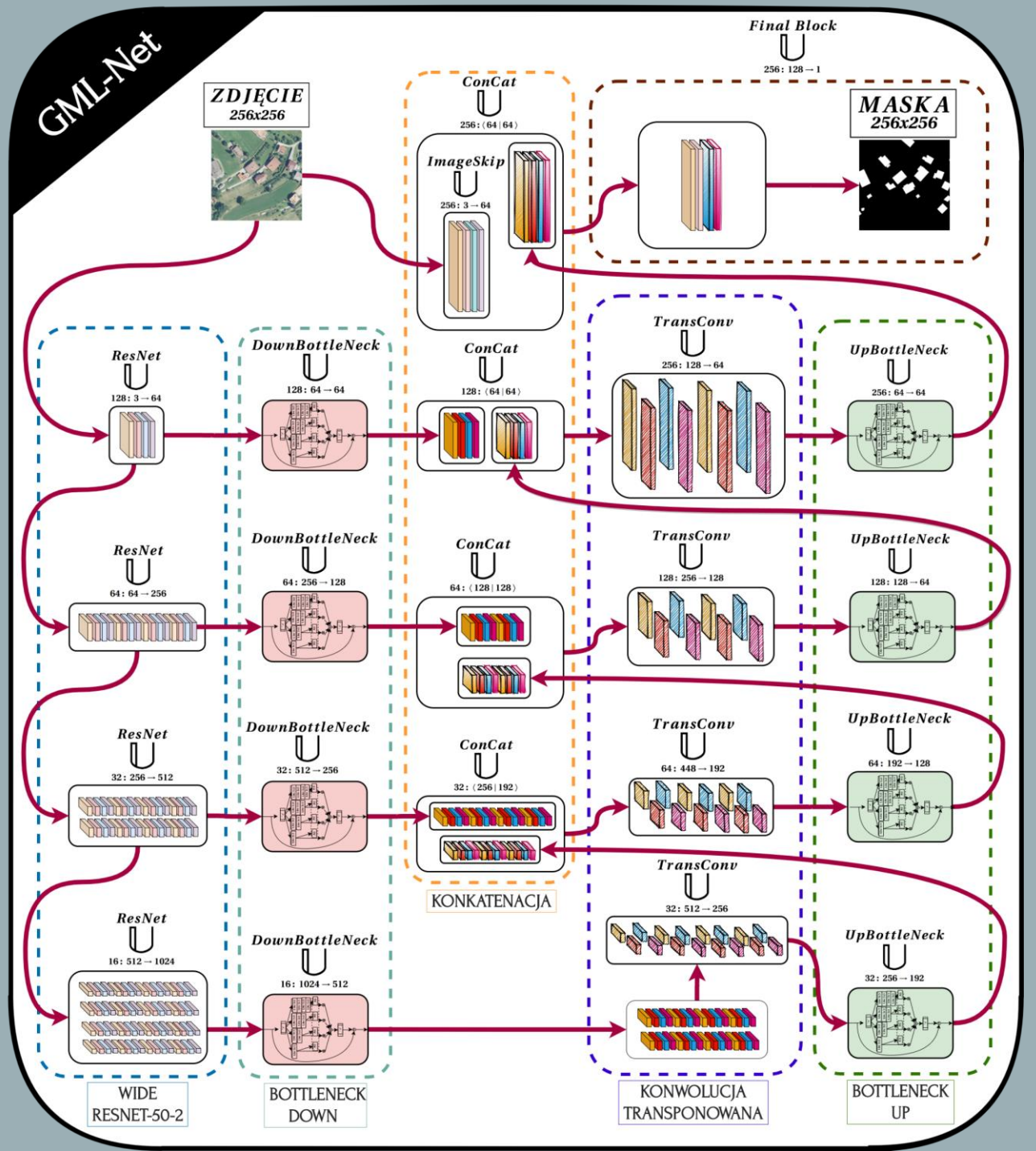
ARCHITEKTURA SIECI *GML-NET*

Architektura sieci *GML-Net* wzorowana jest na architekturze *U-Net*, jest jednak od niej płytsza o jeden poziom ekstrakcji cech / upsamplingu, stąd też występują w niej tylko trzy połączenia rezydualne. Rolę enkodera w sieci *GML-Net* pełni sieć *Wide ResNet-50-2*.

Cechy wyekstraktowane przy pomocy *Wide ResNet-50-2* na danym poziomie rozdzielczości przestrzennej są następnie przetwarzane przez blok *BottleNeck* (wzorowany na blokach sieci OSNet) po to by uzyskać zagregowany zbiór cech z przekroju różnych skal. Zbiór ten jest następnie transferowany, przy pomocy połączeń rezydualnych, bezpośrednio do warstw dekodera.

Rekonstrukcja cech w dekoderze przeprowadzana jest przy pomocy konwolucji transponowanej. Zrekonstruowane cechy są przetwarzane przez blok *BottleNeck*, a następnie przeprowadzana jest ich konkatenacja z cechami uzyskanymi z enkodera - tak połączone mapy cech stają się wsadem do kolejnych, wyższych warstw dekodera.

Sieć *GML-Net* składa się 27 milionów trenowalnych parametrów.


GML-Net
ZDJĘCIE
256x256
ConCat
256: ⟨64|64⟩
ImageSkip
256: 3 → 64
Final Block
256: 128 → 1
MASKA
256x256
ResNet
128: 3 → 64
DownBottleNeck
128: 64 → 64
ConCat
128: ⟨64|64⟩
TransConv
256: 128 → 64
UpBottleNeck
256: 64 → 64
ResNet
64: 64 → 256
DownBottleNeck
64: 256 → 128
ConCat
64: ⟨128|128⟩
TransConv
128: 256 → 128
UpBottleNeck
128: 128 → 64
ResNet
32: 256 → 512
DownBottleNeck
32: 512 → 256
ConCat
32: ⟨256|192⟩
TransConv
64: 448 → 192
UpBottleNeck
64: 192 → 128
ResNet
16: 512 → 1024
DownBottleNeck
16: 1024 → 512
TransConv
32: 512 → 256
UpBottleNeck
32: 256 → 192
KONKATENACJA
WIDE RESNET-50-2
BOTTLENECK DOWN
KONWOLUCJA TRANSPONOWANA
BOTTLENECK UP

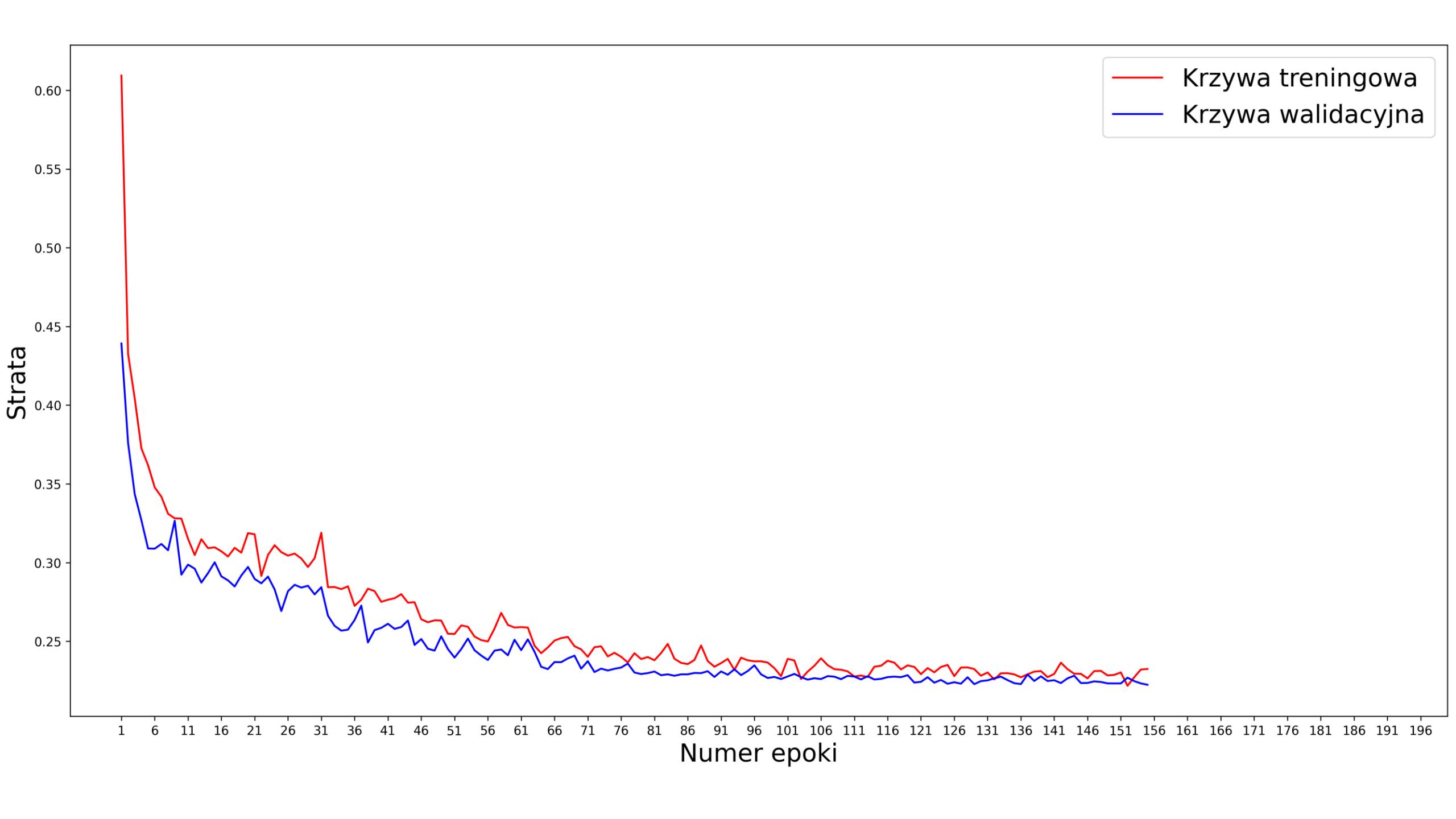
TRENOWANIE SIECI GML-NET

Sieć GML-Net została wytrenowana przy pomocy funkcji straty stanowiącej ważoną sumę: binarnej entropii skrośnej (BCE), Dice Loss (DL) oraz Lovász Hinge Loss (LHL). Po przeprowadzeniu licznych eksperymentów, ustalono iż optymalne wagi poszczególnych części funkcji straty to: 0,3 dla BCE, 0,4 dla DL oraz 0,3 dla LHL:

$$LF = 0{,}3 * BCE + 0{,}4 * DL + 0{,}3 * LHL$$

Po przeprowadzeniu eksperymentów z kilkoma różnymi optymalizatorami oraz po przeanalizowaniu literatury badawczej, jako optymalizator najlepiej realizujący trenowanie sieci *GML-Net* wybrano optymalizator *SGD* z początkową stopą uczenia na poziomie 0,01, momentum o wartości 0,9 i spadkiem wag równym 0,0005. Jako *scheduler* wybrano *ReduceLROnPlateau* ze spadkiem uczenia realizowanym poprzez przemnożenie aktualnej stopy uczenia przez 0,5 przy pięciu epokach bez poprawy stopy uczenia na zbiorze walidacyjnym.

Przez cały okres trenowania sieci *GML-Net* wartość łącznej funkcji straty na zbiorze walidacjnym utrzymywała się poniżej wartości łącznej funkcji straty na zbiorze treningowym. Można to tłumaczyć tym, iż zbiór walidacyjny jest kilkukrotnie mniejszy od zbioru treningowego, może więc charakteryzować się mniejszą zmiennością i stąd generować nieznacznie lepsze wyniki. W czasie treningu nie dochodzi do przetrenowania sieci - zastosowanie silnej augmentacji danych, przyczyniło się do uzyskania wystarczającej odporności modelu na ten problem.


Krzywa treningowa
Krzywa walidacyjna
Strata
Numer epoki
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171 176 181 186 191 196

WYNIKI UZYSKANE NA ZBIORZE WALIDACYJNYM DLA MASEK O ROZDZIELCZOŚCI 256X256

Najlepszy wynik pod kątem łącznej straty na zbiorze walidacyjnym, sieć *GML-Net* uzyskała po 155 epokach obliczeniowych. Średni czas przeliczenia jednej epoki na karcie graficznej Nvidia Tesla V100 wyniósł około 7,5 minuty.

Wartość łącznej funkcji strat dla najlepszej epoki wyniosła 0,2224, a uzyskane metryki były na poziomie: *OA* = 97,08%, *IoU* = 82,42%, *F1S* = 88,22%, *SSIM* = 95,46%.

Uzyskane wartości metryk można uznać za satysfakcjonujące patrząc przez pryzmat wyników uzyskiwanych przez modele zaprezentowane w literaturze dotyczącej detekcji budynków na zdjęciach lotniczych.

Podsumowanie najlepszej epoki uzyskanej przez sieć *GML-Net*

Epoka 155/200

Learning Rate: 4.8828125e-06

Train: BCE Loss: 0.1203, LH Loss: 0.4504, Dice Loss: 0.1532, Final Loss: 0.2325
OA: 96.88%, IoU: 78.87%, F1S: 84.68%, SSIM: 95.31%

Valid: BCE Loss: 0.1165, LH Loss: 0.4679, Dice Loss: 0.1178, Final Loss: 0.2224
OA: 97.08%, IoU: 82.42%, F1S: 88.22%, SSIM: 95.46%

Najlepszy jak dotad model!
Czas przeliczenia biezacej epoki: 7m 31s

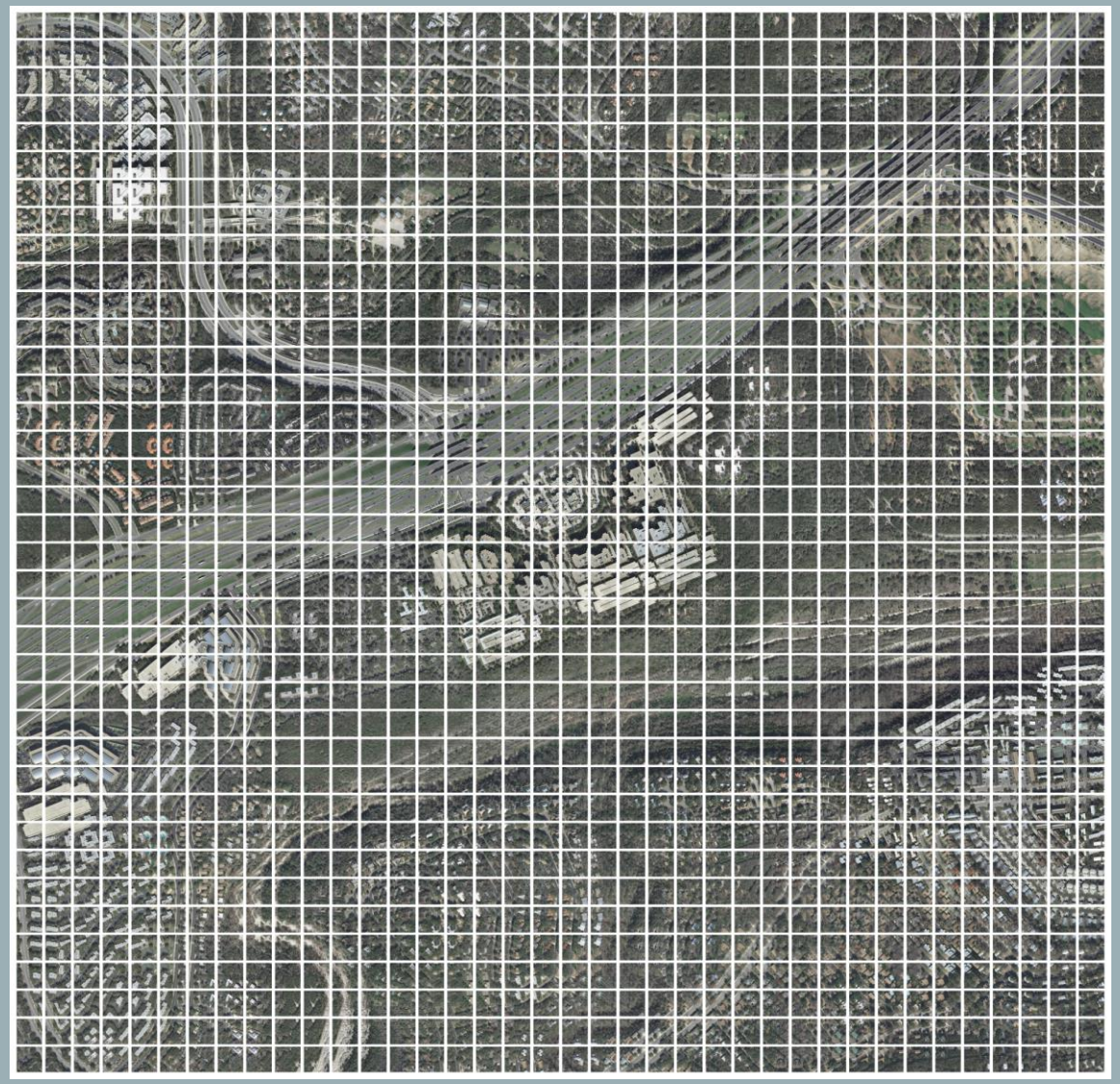
GENEROWANIE PREDYKCJI DLA 1444 PODPRÓBEK ZDJĘĆ WALIDACYJNYCH

W związku z tym, iż celem omawianego w bieżącej pracy badania własnego było skonstruowanie głębokiej sieci neuronowej, która umożliwi efektywną detekcję zabudowań na zdjęciach lotniczych o wysokiej rozdzielczości, konieczne stało się wypracowanie procesu, który umożliwi predykcję masek o rozdzielczości 5000x5000.

Aby to zrobić postanowiono wygenerować predykcje dla podpróbek o rozdzielczości 256x256 a następnie te podpróbki miały zostać połączone w jedną łączną maskę o rozdzielczości 5000x5000.

Każde zdjęcie zostało podzielone na 1444 nakładające się podpróbki, szerokość nakładania się ustalono na połowę wymiaru podpróbki. Każda podpróbka została przepuszczona przez sieć *GML-Net* w celu wygenerowania dla niej maski.

Przykładowe zdjęcie rozbite na 1444 podpróbki



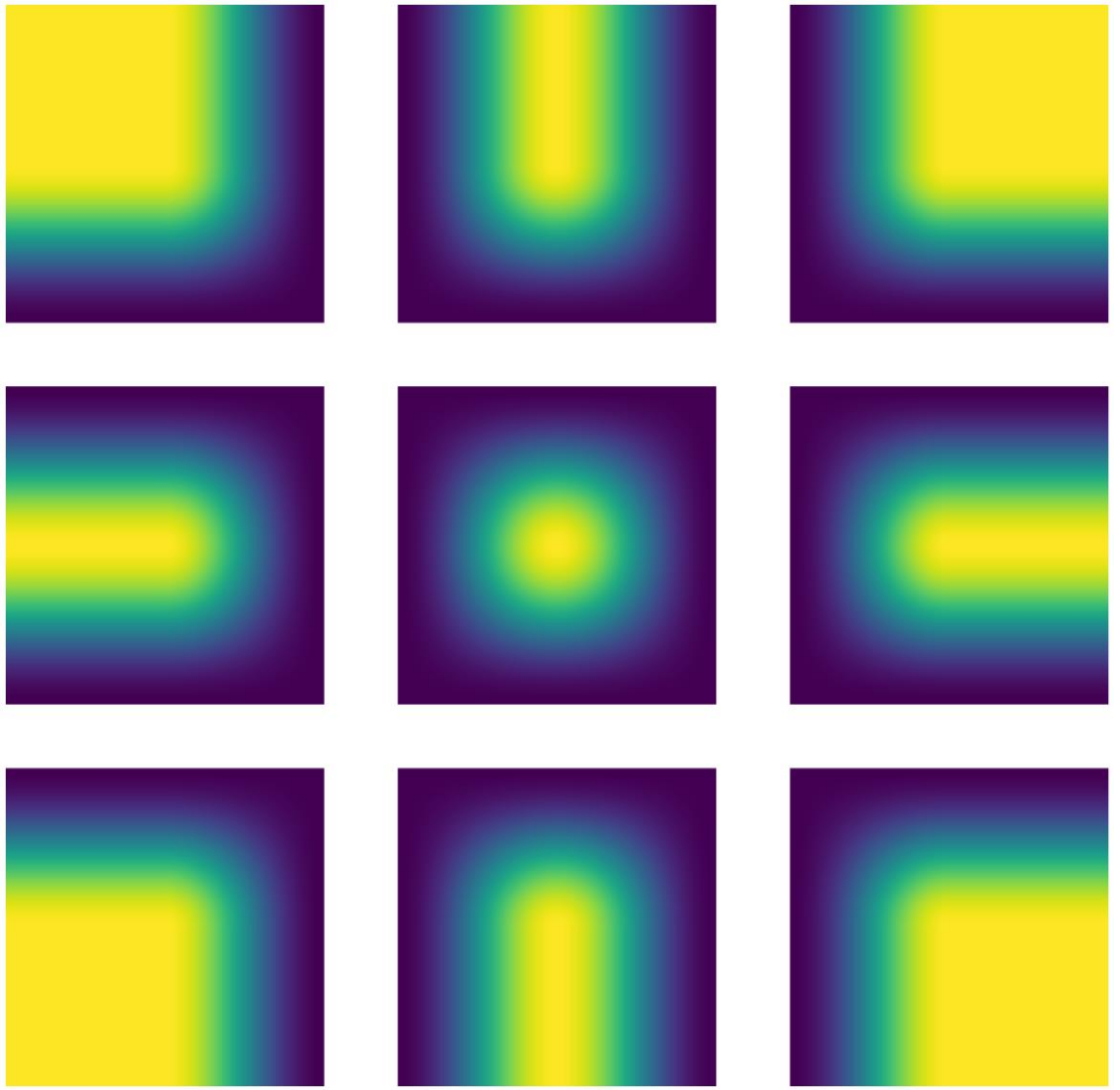
ŁĄCZENIE 1444 PREDYKCJI W JEDNĄ ŁĄCZNĄ MASKĘ 5000X5000

Aby móc efektywnie połączyć uzyskane 1444 predykcje w jedną łączną maskę o rozdzielczości 5000x5000 postanowiono skorzystać z dwuwymiarowych okien Hanna.

Predykcje uzyskane dla poszczególnych podpróbek były przemnażane przez dwuwymiarowe okna Hanna, stanowiące *de facto* macierze wag, które główny nacisk kładły na predykcje znajdujące się w środku podpróbki (bazowa wersja okna Hanna 2D) - w ten sposób redukowany był efekt złych predykcji krawędzi.

Dla sytuacji, w których predykcja pochodziła z podpróbki znajdującej się na krawędzi zdjęcia walidacyjnego zdefiniowano 8 dodatkowych wariantów okna Hanna 2D, tak żeby w takich przypadkach nacisk został położony na predykcje znajdujące się na krawędziach podpróbki.

Wizualizacja dziewięciu okien Hanna 2D służących do właściwego ważenia predykcji poszczególnych fragmentów zdjęć walidacyjnych



WYNIKI UZYSKANE NA ZBIORZE WALIDACYJNYM DLA MASEK O ROZDZIELCZOŚCI 5000X5000

Sieć *GML-Net* osiągnęła zadowalającą finalną skuteczność na zbiorze walidacyjnym przy predykcji masek o rozmiarach 5000x5000 - uzyskano następujące finalne wartości metryk: *OA* = 96,44%, *IoU* = 75,97%, *F1S* = 86,07% oraz *SSIM* = 94,55%.

Wyniki te zostały uzyskane przy średnim czasie generowania predykcji jednej maski 256x256 na poziomie 0,0003 sekundy oraz średnim czasie generowania predykcji maski łącznego obrazka 5000x5000 na poziomie 15 sekund.

Oryginalna maska (*ground truth*)

Maska wygenerowana przez sieć GML-Net

PORÓWNANIE METRYK JAKOŚCI PREDYKCJI PODPRÓBKI 256X256 VS. ZDJĘCIA 5000X5000

Rodzaj metryki	Wyniki dla masek 256x256	Wyniki dla masek 5000x5000
OA	97,08%	96,44%
IoU	82,42%	75,97%
F1S	88,22%	86,07%
SSIM	95,46%	94,55%

Jak można się było spodziewać metryki jakości predykcji modelu wyliczane na poziomie pełnych zdjęć o rozdzielczości 5000x5000 są gorsze od metryk wyliczanych na poziomie podpróbek.

Ogólna dokładność pogarsza się o 0,64 punktu procentowego, współczynnik podobieństwa Jaccarda spada aż o 6,45 p.p., wynik F1 jest niższy o 2,35 p.p. a indeks podobieństwa strukturalnego jest gorszy o 0,91 p.p.

WYNIKI UZYSKANE NA ZBIORZE TESTOWYM DLA MASEK O ROZDZIELCZOŚCI 5000X5000

Miasto	OA	IoU
Bellingham	97,11%	71,39%
Bloomington	97,34%	71,83%
Innsbruck	96,99%	74,89%
San Francisco	91,72%	74,98%
Tyrol Wschodni	98,01%	77,85%
Łącznie:	**96,23%**	**74,42%**

W związku z tym, iż autorzy zbioru *Inria Aerial Image Labeling Dataset* nie udostępnili masek *ground truth* dla zbioru testowego, jedynym sposobem by poznać skuteczność sieci *GML-Net* na zbiorze testowym było wysłanie im predykcji 180 masek dla tego zbioru, aby to oni wyliczyli finalną skuteczność stworzonej sieci.

W tabeli powyżej zaprezentowane są wyniki sieci *GML-Net n*a zbiorze testowym *IAILD* wyliczone przez autorów tego zbioru. Jak widać omawiana w bieżącym rozdziale sieć osiągnęła łączną skuteczność mierzoną przy pomocy metryki *OA* na poziomie 96,23% a mierzoną przy pomocy metryki IoU na poziomie 74,42%. Najlepsze wyniki sieć *GML-Net* uzyskała dla miasta Tyrol Wschodni, a najgorsze dla miast San Francisco (pod kątem OA) oraz Bellingham (pod kątem IoU).

UZYSKANE WYNIKI NA TLE LITERATURY BADAWCZEJ

Porównując uzyskane wyniki do wyników zaprezentowanych w przeglądzie literatury można śmiało stwierdzić, iż sieć *GML-Net* jest w stanie generować wyniki o zbliżonej jakości do modeli przedstawionych w literaturze, odstając o zaledwie niecały punkt procentowy od najlepszego wyniku pod kątem metryki *OA* i o niecałe sześć punktów procentowych pod kątem metryki *IoU*,

Wyniki uzyskane przez sieć *GML-Net* można również porównać do wyników prezentowanych przez autorów zbioru *IAILD* na ich stronie internetowej. Znajduje się tam tabela przedstawiająca wartości uzyskiwanych metryk *OA* i *IoU* dla 110 modeli, których autorzy zgodzili się na publikację wskaźników jakości predykcji ich sieci. Średnia wartość ogólnej dokładności dla tych 110 modeli wynosi 95,46%, a średnia wartość indeksu Jaccarda wynosi 70,02%, co oznacza, iż sieć *GML-Net* uzyskuje wyniki istotnie wyższe od średnich dla tego zbioru - mówiąc dokładniej, zajmuje 29. miejsce pod kątem *OA* oraz 30 pod kątem *IoU*.

Podsumowanie wyników uzyskanych na zbiorze IAILD przez autorów innych badań zajmujących się problematyką detekcji budynków na zdjęciach lotniczych.

Tytuł artykułu	Uzyskane wyniki
Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks [1]	OA: 95.17% IoU: 70.14%
Semantic Segmentation from Remote Sensor Data and the Exploitation of Latent Learning for Classification of Auxiliary Tasks [2]	OA: 97.14% IoU: 80.32%
Building Footprint Generation by Integrating Convolution Neural Network with Feature Pairwise Conditional Random Field (FPCRF) [37]	OA: 95.81% F1S: 87.65% IoU: 74.79%
Polygonal Building Segmentation by Frame Field Learning [7]	IoU: 78.00%

PODSUMOWANIE

- Zaprezentowana w bieżącej prezentacji sieć GML-Net pozwoliła na uzyskanie zadowalającej skuteczności w detekcji budynków na zdjęciach lotniczych pochodzących ze zbioru *Inria Aerial Image Labeling Dataset*.

- Nie udało się w prawdzie osiągnąć lepszych rezultatów niż aktualne wyniki *state of art* zaprezentowane w pracy *Semantic Segmentation from Remote Sensor Data and the Exploitation of Latent Learning for Classification of Auxiliary Tasks*, ale mimo to wyniki uzyskane przy pomocy sieci *GML-Net* można uznać za satysfakcjonujące.

- Za duże pole do dalszego rozwoju uzyskanej sieci można uznać sposób łączenia predykcji masek o rozdzielczości 256x256 w jedną łączną maskę o rozdzielczości 5000x5000, gdyż w tym procesie skuteczność predykcji istotnie się pogarszała, szczególnie patrząc przez pryzmat metryki *Intersection over Union*.

- Za duże zalety sieci GML-Net można uznać jej ciekawą architekturę wykorzystująca elementy sieci *ResNet*, *U-Net* oraz *ICT-Net*, a także nowatorską funkcję straty stanowiąca ważoną sumę *Binary Cross-Entropy Loss*, *Dice Loss* oraz *Lovász hinge loss*.