Zielona Góra, 22 lutego 2024 r.

prof. dr hab. inż. Dariusz Uciński
Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

RECENZJA
rozprawy doktorskiej Pana mgr inż. Daniela Dworaka
pt. ***Niskopoziomowa fuzja danych sensorycznych do detekcji obiektów w systemie percepcji pojazdu autonomicznego bazująca na technikach uczenia maszynowego***
opracowana na wniosek Rady Dyscypliny Naukowej Automatyka, Elektronika i Technologie Kosmiczne Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie

I. Obszar problemowy rozprawy

Jednym z najbardziej perspektywicznych i najciekawszych kierunków rozwoju automatyki jest jazda autonomiczna, odnosząca się do pojazdów poruszających się bez interwencji kierowcy. Wprawdzie samochody dostępne obecnie na rynku nie są jeszcze w stanie działać w pełni autonomicznie, czyli zupełnie bez interwencji człowieka, jednak intensywne badania naukowe i przemysłowe przekładają się na gwałtowny postęp i pokonywanie kolejnych etapów automatyzacji. Jazda autonomiczna niesie ogromny potencjał łagodzenia skutków dużego natężenia ruchu drogowego i poprawy bezpieczeństwa. Percepcja otoczenia, obejmującego infrastrukturę uliczną, inne pojazdy, pieszych, sygnalizację świetlną i znaki drogowe, stanowi punkt wyjścia do monitorowania i tworzenia jego trójwymiarowej mapy. W tym celu wykorzystuje się teledetekcję, przede w oparciu o radary, GPS, kamery i lidary. Komputer pokładowy ma w czasie rzeczywistym przetwarzać napływające dane pomiarowe i podejmować decyzje dotyczące działania pojazdu, planując jego trasę i kontrolując manewry.

Naturalnym wymogiem stawianym systemowi percepcji jest jego maksymalnie duża dokładność oraz odporność na niepewności i zakłócenia. Narzucającym się środkiem jego spełnienia jest integracja różnych czujników, pozwalająca wykorzystać komplementarność i redundantność ich charakterystyk. Powstaje jednak przy tym problem, które czujniki wybrać oraz jak dokonać fuzji otrzymywanych z nich danych, posiadających często zupełnie różne właściwości. Do tej pory stanowi on ogromne wyzwanie, stając się przedmiotem bardzo intensywnych badań naukowych wielu zespołów na świecie.

Prototypy pojazdów autonomicznych szeroko stosują kamery do wykrywania obiektów, segmentacji i śledzenia. Lidary określają odległość od otaczających obiektów poprzez ich oświetlanie impulsami światła laserowego i pomiar czasu powrotu wiązki światła oraz zmiany długości fali, na podstawie czego

można utworzyć trójwymiarowy model otoczenia. Komplementarne funkcjonalności kamer i lidarów sprawiły, że fuzja danych z nich otrzymywanych stała się nie tylko bardzo modnym tematem badawczym, ale również doprowadziła do wyjątkowo wysokiej dokładności detekcji obiektów w scenach dwu- i trójwymiarowych. Pomimo ich mocnych stron, zarówno lidary, jak i kamery dzielą te samą wadę, jaką jest wrażliwość na niekorzystne warunki atmosferyczne (np. deszcz, mgłę, śnieg), które mogą znacznie zmniejszyć ich pole widzenia oraz możliwości rozpoznawania obiektów. Co więcej, wysoki koszt lidarów nadal stanowi przeszkodę w ich szerokim rozprzestrzenieniu.

W porównaniu z lidarami i kamerami, radary wykazują zdecydowanie wyższą skuteczność w trudnych warunkach oświetleniowych i pogodowych. Mogą również zapewnić dokładne oszacowania prędkości dla wszystkich wykrytych obiektów w oparciu o efekt Dopplera. Szeroko stosuje się je w zaawansowanych systemach wspomagania kierowcy (ADAS), tempomatach adaptacyjnych (ACC), układach asystenta zmiany pasa ruchu (LCA), czy też systemach autonomicznego hamowania awaryjnego (AEB). Niestety, mimo tych sukcesów, niewiele badań koncentruje się na fuzji danych z radarów i kamer. Jednym z powodów są ograniczenia danych wyjściowych radaru, takie jak niska rozdzielczość, rzadkie chmury punktów, niepewność określenia kąta podniesienia i odbicia zakłócające od innych obiektów. Innym powodem jest to, że dostępne zbiory danych zawierające zarówno dane z radarów, jak i z kamer do zastosowań w pojazdach autonomicznych są stosunkowo nieliczne, co sprawia, że przeprowadzenie dogłębnej analizy jest nadal bardzo trudne. Ponadto, zastosowanie lub adaptacja istniejących algorytmów dedykowanych lidarom do radarowych chmur punktów daje raczej słabe wyniki z powodu dużych różnic między chmurami punktów generowanych przez lidary i radary: te drugie są znacznie rzadsze. W konsekwencji, fuzję danych z radaru i kamery nadal uważa się za duże wyzwanie, chociaż przewiduje się, że rola tego sposobu obserwacji otoczenia będzie systematycznie rosnąć (stwierdzają to np. S. Yao i in. w obszernej pracy przeglądowej pt. *Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review*, niedawno zaakceptowanej w IEEE Transactions on Intelligent Vehicles).

Właśnie w tym kontekście recenzowana praca Pana mgr inż. Daniela Dworaka, poświęcona w całości metodom niskopoziomowej fuzji danych dla par czujników kamera-radar oraz kamera-lidar, wykorzystującym głębokie sieci neuronowe, jest pozycją niezwykle ambitną i aktualną. Zasadniczo, oryginalny pomysł Autora polega na dopasowaniu przestrzeni danych obu czujników (z powodu zupełnie odmiennych orientacji próbek względem wspólnego układu współrzędnych pojazdu) oraz zaproponowaniu trzech wariantów wieloskalowej fuzji danych i detekcji obiektów dokonywanych przez sieć głęboką sieć neuronową o architekturze wzorowanej na modelu EfficientDet. Istotnym wątkiem rozprawy jest również wizualna objaśnialność detekcji dokonywanych przez stosowane modele sieci splotowych, odpowiadająca najnowszym trendom uczenia maszynowego, która doprowadziła do zaproponowania wariantu określanego jako fuzja agregacyjna oparta o zasięg (*ang.* range-based aggregation fusion), będącego najbardziej wartościowym i najbardziej nowatorskim wynikiem Autora, prowadzącym do znakomitych rezultatów detekcji obiektów w przypadku zastosowania kamery i radaru, o jakości porównywalnej z zastosowaniem o wiele droższego lidaru. Co więcej, obszerną część rozprawy zajmuje praktyczna weryfikacja działania zaproponowanych metod w oparciu o nietrywialne zadania testowe (zestawy danych KITTI i NuScenes).

Biorąc pod uwagę wszystkie wymienione czynniki, sformułowane na str. 3 cele pracy, jak również wynikające z nich zrealizowane zadania szczegółowe, są jasne i dobrze określone. Sprowadzają się one do wykazania, że niskopoziomowa fuzja danych napływających z kamery i radaru lub z kamery i lidaru, przeprowadzana w ramach jednej architektury głębokiej sieci dokonującej jednocześnie detekcji i klasyfikacji obiektów, znacząco poprawia dokładność percepcji i redukuje niepewności, prowadząc do lepszego rozpoznania środowiska niż w przypadku pojedynczego czujnika. Tak zarysowaną problematykę rozprawy uważam za istotną i nadzwyczaj aktualną, o rezultatach mogących otworzyć nowy nurt badań nad niskopoziomową fuzją danych na potrzeby detekcji obiektów i segmentacji semantycznej w jeździe pojazdów autonomicznych. Fakt ten przesądza o pozytywnej ocenie wybranego tematu jako przedmiotu opiniowanej rozprawy doktorskiej.

II. Koncepcja oraz realizacja rozprawy

Obszerna rozprawa, napisana w języku angielskim i licząca 135 stron numerowanych, składa się ze wstępu, trzech rozdziałów wprowadzających czytelnika do zagadnień jazdy autonomicznej, uczenia głębokiego w systemach percepcji i miar oceny jakości metod detekcji, jednego zasadniczego rozdziału przedstawiającego koncepcję trzech proponowanych technik fuzji danych, trzech rozdziałów raportujących wyniki eksperymentów weryfikujących działanie tych metod w praktyce, rozdziału opisującego rezultaty analizy wykorzystującej techniki objaśnialnej sztucznej inteligencji, oraz rozdziału podsumowujących uwag końcowych. Załączony niezwykle obszerny wykaz 156 pozycji cytowanej literatury bardzo dobrze odzwierciedla stan badań w zakresie tematycznym rozprawy.

Pracę rozpoczyna *Wstęp*, na które składa się przedstawienie motywacji zagadnień rozprawy, cele pracy (w tym momencie już intuicyjnie jasne), a także zwięzłe zestawienie osiągniętych rezultatów oraz dokumentujących je publikacji i zgłoszeń patentowych. Rozdział kończy charakterystyka struktury rozprawy.

Rozdział 2 stanowi bardzo udane wprowadzenie do zagadnień jazdy autonomicznej. Charakteryzuje się tu sześć poziomów autonomiczności pojazdów zdefiniowanych przez Society of Automotive Engineers oraz dokonuje analizy porównawczej czujników stosowanych w pojazdach autonomicznych (kamery, lidary, radary). Rozdział uzupełnia wprowadzenie do wysoko- i niskopoziomowej fuzji danych.

Głębokie sieci neuronowe, a zwłaszcza sieci splotowe, dokonały bezprecedensowego przełomu w dziedzinie wizji komputerowej i rozpoznawania obrazów, począwszy od klasyfikacji obrazów do detekcji obiektów i segmentacji semantycznej. Nic więc dziwnego, że praktycznie zdominowały współczesne techniki fuzji danych otrzymywanych z czujników rozważanych w rozprawie. *Rozdział 3* stanowi przegląd podejść wykorzystujących je w systemach percepcji pojazdów autonomicznych najpierw opartych o dane wyłącznie z kamery, lidaru lub radaru, a następnie w kontekście niskopoziomowej fuzji danych (podejścia jednowidokowe i wielowidokowe)

Rozdział 4 podaje zestaw metryk stosowanych do oceny jakości systemów percepcji wraz z porównaniem ich przydatności w rozważanych zadaniach. Wymienia się tu miary podobieństwa między zaetykietowanymi i przewidywanymi ramkami ograniczającymi (stosunek objętości iloczynu do pola sumy, odległość euklidesowa między środkami oby ramek), miary jakości klasyfikacji (macierz pomyłek, czułość, precyzja, współczynnik F1), metryki wykrywania obiektów (np. średnia arytmetyczna precyzji, średnia arytmetyczna błędu przesunięcia, średnia arytmetyczna błędu orientacji). Rozdział uzupełnia krótkie wprowadzenie do dziedziny objaśnialnej sztucznej inteligencji, z grubsza sprowadzającego się tu do budowy map cieplnych wskazujących części obrazu wejściowego najbardziej wpływających na predykcję modelu.

Po przeczytaniu pierwszych czterech rozdziałów czytelnik posiada dobrą ogólną orientację w zakresie omawianych zagadnień i dotychczas stosowanych podejściach.

Rozdział 5 jest najważniejszym rozdziałem pracy, prezentując oryginalne rozwiązania zaproponowane przez Autora. Określa się je zbiorczo międzydziedzinowym dopasowywaniem przestrzennym (*ang.* cross-domain spatial matching method, CDSM), chociaż ta nazwa oddaje tylko jeden aspekt proponowanej techniki, związany z transformacją dwuwymiarowego obrazu pochodzącego z kamery i trójwymiarowej chmury punktów generowanych przez radar lub lidar. Zazwyczaj tę drugą rzutuje się do dwuwymiarowego układu współrzędnych kamery, jednak w rozprawie zaproponowano inne podejście, polegające na fuzji surowych map cech, zbudowanych przez odpowiednie architektury sieci splotowych. Dla kamery jest to bardzo efektywna (również pamięciowo) znana sieć EfficientDet. Jej podsieć szkieletową stanowi sieć EfficientNet2, której trzy z pięciu poziomów generują cechy odpowiadające różnym rozdzielczościom, odpowiadającym z kolei różnym skalom i poziomom abstrakcji. Wyjścia tej podsieci przetwarza dwukierunkowa podsieć piramidy cech (*ang.* Bi-directional

Feature Pyramid Network, BiFPN), przekształcając je w mapy cech obrazu. BiFPN wprowadza dodatkowe wagi, które dostrajają się do ważności różnych cech. Podobne mapy cech chmury punktów buduje sieć dedykowana radarowi/lidarowi. Różnica w jej architekturze polega na wykorzystaniu w charakterze podsieci szkieletowej struktury agregacji warstw głębokich (*ang.* deep layer aggregation), której wejścia generuje moduł ekstraktora cech wokseli. Każdy z trzech poziomów szczegółowości skali przestrzennej zawiera 256 map cech (interpretowanych jako kanały). W rozprawie proponuje się trzy nowe techniki fuzji map cech zbudowanych przez sieć dla kamery oraz sieć dla radaru/lidaru:

- Fuzja jeden-do-jednego (*ang.* one-to-one fusion): polega na prostej konkatenację kanałów na danym poziomie;

- Fuzja agregacji cech (*ang.* feature-wise aggregation fusion): przed konkatenacją kanałów na danym poziomie przeprowadza się agregację map cech kamery wzdłuż osi pionowej układu współrzędnych pojazdu, połączoną z tzw. uszczegółowieniem (*ang.* refinement), sprowadzającym się do działania filtrów splotowych określających korelację przestrzenną między cechami.

- Fuzja agregacji opartej o zasięg (ang. range-based aggregation fusion): dla kamery tworzy się mapy cech na pięciu poziomach, z których każdy odpowiada jednemu z pięciu zakresów odległości od pojazdu; przed konkatenacją kanałów na danym poziomie przeprowadza się agregację map cech kamery względem osi układu współrzędnych skierowanej w kierunku ruchu pojazdu, połączoną z działaniem filtrów splotowych.

Mapy cech połączone w wyniku fuzji przetwarza sieć BiFPN połączona z podsiecią dokonującą predykcji (detekcji i klasyfikacji obiektów).

Należy tu również zwrócić uwagę na bardzo ciekawy autorski pomysł wykorzystania architektury EfficientDet w połączeniu z modułem CDSM do detekcji obiektów w przestrzeni trójwymiarowej wyłącznie za pomocą kamery.

W kolejnych czterech rozdziałach Autor przedstawia wyniki eksperymentów potwierdzających efektywność proponowanych przez Niego metod. Ta część stanowi aż połowę objętości pracy, jednak jest równie interesująca, jak część pierwsza. Duże uznanie budzi przede wszystkim ogromny nakład pracy włożony w przeprowadzenie wszystkich badań, tak bardzo dalekich od trywialności. Ich rezultaty potwierdziły w praktyce zasadność zaproponowanej metodologii.

Podejścia proponowane w pracach badawczych dotyczących fuzji danych czujników na potrzeby systemów percepcji pojazdów autonomicznych testuje się zazwyczaj w oparciu o kilka ogólnie dostępnych zestawach danych. W pracy wykorzystano zestawy KITTI (kamera + lidar) oraz NuScenes (6 kamer + lidar + 5 radarów), które wydają się być najczęściej eksploatowanymi w literaturze dotyczącej rozważanej fuzji. Ich szczegółowy opis zawiera *rozdział 6*. Uzupełniają go szczegóły procesu uczenia (funkcję strat, będącą sumą członów odpowiadających klasyfikacji i regresji, minimalizuje się z zastosowaniem algorytmu ADAM; opis zawiera również sposób strojenia hiperparametrów).

Rozdziały 7 i 8 dotyczą wyników eksperymentów obliczeniowych (pierwszy z nich – dla proponowanych architektur sieci głębokich zastosowanych dla pojedynczych czujników, drugi – dla proponowanych technik niskopoziomowej fuzji danych). Uwagę zwraca bardzo duża staranność w prezentowaniu rezultatów (bardzo czytelne wykresy i tabele, szczegółowe omówienia). Fuzja danych kamery i lidaru prowadzi do poprawy względem samodzielnych czujników, jednak nie jest ona spektakularna. Dużo bardziej perspektywiczne jest połączenie kamery i radaru. Właśnie taka kombinacja podlegała dalszej optymalizacji, prowadząc do bardzo dobrych wyników, zbliżonych do tych otrzymywanych z zastosowaniem znacznie droższego lidaru. Co więcej, gdy jeden z czujników nie

wykrywa obiektu, robi to drugi, co dodatkowo potwierdza zasadność zaproponowanego podejścia.

Rozdział 9 zawiera wyniki zaadoptowania znanej techniki Grad-CAM, należącej do gwałtownie rozwijającego się nurtu objaśnialnej sztucznej inteligencji (*ang.* explainable artificial intelligence), stosowanej do wizualnego objaśniania klasyfikacji dokonywanych przez sieci głębokie. Technika prowadzi do map cieplnych, na których intensywniej odzwierciedla się obszary obrazu najbardziej wpływające na przynależność do danej klasy. Wprawdzie wyniki tu prezentowane mają charakter wstępny i odnoszą się do pojedynczych czujników (kamera i lidar), jednak dla nietrywialnych architektur sieci głębokich zaproponowanych w rozprawie wymagały pewnych przeformułowań oryginalnej techniki Grad-CAM. Są na tyle obiecujące, że ten kierunek należy uznać za bardzo perspektywiczny.

Rozprawę kończy podsumowanie oryginalnych wyników naukowych oraz charakterystyka otwartych problemów badawczych.

Oceniając merytorycznie całą rozprawę stwierdzam, że jest ona napisana na bardzo dobrym poziomie. Zawiera jasno sformułowany i ważny problem naukowy, oraz prezentuje poprawne rozwiązanie tego problemu, które zostało uzyskane przez Autora samodzielnie i z zastosowaniem właściwej metodologii naukowej. Na podstawie przedstawionego skrótowo omówienia treści całej rozprawy doktorskiej należy odnotować, że jej Autor wykazał się dobrymi umiejętnościami formułowania problemów naukowo-badawczych oraz ich efektywnego rozwiązywania z zastosowaniem zaawansowanych narzędzi uczenia maszynowego, wizji komputerowej, robotyki mobilnej, rozpoznawania obrazów i technik algorytmicznych. Już na podstawie wstępnej analizy można stwierdzić, że rozprawa stanowi dzieło wartościowe, zdecydowanie odpowiadające wymaganiom stawianym przez stosowne przepisy.

Pod względem redakcyjnym praca napisana jest z bardzo dużą dbałością o szczegóły. Użyte słownictwo odpowiada powszechnie stosowanemu. Jak na tak dużą objętość, praktycznie nie zawiera błędów składu, co tym bardziej koresponduje z jej bardzo dobrym poziomem merytorycznym. Na szczególne podkreślenie zasługuje to, że pracę napisano w bardzo dobrym języku angielskim.

III. Oryginalne osiągnięcia

Chociaż fuzja danych sensorycznych do detekcji obiektów w systemach percepcji pojazdów autonomicznych cieszy się w okresie ostatnich kilku lat coraz większym zainteresowaniem, nie tylko ze względu na jej motywacje praktyczne, ale również jako niezmiernie interesujące pole zastosowań nowoczesnego uczenia maszynowego, to jednak poszukiwanie sposobów uczynienia go jeszcze bardziej efektywnym, a w szczególności przystosowania go do niskobudżetowych czujników, takich jak kamera i radar, jest zadaniem nadzwyczaj trudnym i wciąż aktualnym. Przedstawiony w pracy opis problemu, jego analizę oraz zaproponowane metody i algorytmy obliczeniowe uważam za najważniejszy wkład Autora w rozważaną dziedzinę. Główną zaletą podejścia proponowanego w rozprawie, bardzo dalekiego od trywialności, jest wysoka efektywność obliczeniowa oraz względnie prosta możliwość jego rozwinięcia do działającego prototypu.

Przyjmując, że głównym celem rozprawy było pokazanie, że umiejętnie przeprowadzona niskopoziomowa fuzja danych napływających z kamery i radaru lub z kamery i lidaru w ramach jednej architektury sieci głębokiej, dokonującej jednocześnie detekcji i klasyfikacji obiektów, znacząco poprawia dokładność percepcji i redukuje niepewności, prowadząc do lepszego rozpoznania środowiska niż w przypadku pojedynczego czujnika, należy stwierdzić, że cel ten Autor osiągnął. Co więcej, weryfikacji rezultatów dokonano w oparciu o uznane benchmarki dobrze odzwierciedlające warunki rzeczywistego użytkowania.

W szczególności, za najważniejsze rezultaty rozprawy uważam następujące:

1. zaproponowanie zupełnie nowego podejścia do niskopoziomowej fuzji danych, określonego jako międzydziedzinowe dopasowywanie przestrzenne (*ang.* cross-domain spatial matching method, CDSM), polegającego na ujednoliceniu odczytów par czujników (kamera-radar albo kamera-lidar) poprzez wyrażenie ich we wspólnym trójwymiarowym układzie współrzędnych pojazdu oraz zastosowanie jednej z trzech nowych strategii fuzji zintegrowanej z architekturą głębokiej sieci neuronowej, wykorzystujących wieloskalowe mapy cech obrazów i chmur punktów, budowane automatycznie przez podsieci splotowe;

2. zaproponowanie nowatorskiego podejścia do detekcji obiektów w przestrzeni trójwymiarowej przez pojedynczą kamerę, opartego o integrację modułu CDSM z architekturą sieci EfficientDet;

3. nietrywialna adaptacja techniki Grad-CAM, jednej z metod objaśnialnej sztucznej inteligencji, do określania cech i części obrazów najbardziej wpływających na detekcję i klasyfikację obiektów, nowatorskiej zwłaszcza dla danych pochodzących z lidaru;

4. pozytywna weryfikacja proponowanych rozwiązań w oparciu o dane rzeczywiste pochodzące z benchmarków KITTI i NuScenes.

Należy podkreślić, że uzyskane rezultaty są udokumentowane publikacjami zarówno w czasopismach (*Energies*, MDPI, IF = 3.2, 140 pkt wg MNiSW), jak i na trzech konferencjach międzynarodowych (*Methods and Models in Automation and Robotics*, dwukrotnie *Polish Control Conference*). Uzupełniają to cztery zgłoszenia patentowe, co jest godne dużego uznania.

W podsumowaniu, należy stwierdzić, że **sformułowany cel rozprawy został osiągnięty, a jej Autor wykazał się głęboką wiedzą i umiejętnościami niezbędnymi do samodzielnego rozwiązywania problemów naukowo-technicznych w dyscyplinie *automatyka, elektronika, elektrotechnika i technologie kosmiczne*.**

IV. Uwagi i komentarze

Przedstawiona do recenzji praca zawiera istotną treść naukową i wiele nowych wyników. Stanowi logiczną całość począwszy od praktycznego uzasadnienia problemu, poprzez jego formalizację, aż do rozwiązania różnorakich wersji problemu z przykładami zastosowań zaproponowanej metodologii w nietrywialnych zadaniach testowych. Praca prezentuje wysoki poziom naukowy, a Autor biegle posługuje się nowoczesnymi narzędziami uczenia maszynowego, wizji komputerowej, robotyki mobilnej, rozpoznawania obrazów i technik algorytmicznych.

Lektura rozprawy skłania jednak również do sformułowania następujących uwag krytycznych:

1. Rezultaty zademonstrowane w rozprawie dowodzą tego, że zaproponowane podejście do niskopoziomowej wieloczujnikowej fuzji danych może znacząco poprawić jakość detekcji w porównaniu z wykorzystaniem pojedynczych czujników. Nieco jednak szkoda, że nie dokonano próby porównania jakości proponowanego sposobu fuzji z alternatywnymi technikami znanymi z literatury (np. z tymi opisanymi w rozdz. 3.3). Być może w literaturze dałoby się odnaleźć wartości wskaźników jakości detekcji osiągane przez takie alternatywy na identycznych zestawach danych i potraktować je jako wartości odniesienia w celu obiektywniejszej oceny możliwości zaproponowanego podejścia. Pomocna wydaje się tu np. praca He W, Deng Z, Ye Y, Pan P. ConCs-Fusion: A Context Clustering-Based Radar and Camera Fusion for Three-Dimensional Object Detection. *Remote Sensing*. 2023; 15(21):5130, dotycząca bardzo

podobnego zagadnienia, również testowanego na zestawie NuScenes.

2. Ciekawym byłoby sprawdzenie na ile zaproponowane podejście jest odporne na uszkodzenie jednego z czujników (w najprostszej sytuacji po prostu jego wyłączenie). Wprawdzie w rozdz. 8.4 omawia się sytuacje, w których albo kamera, albo radar nie wykrywają obiektu, a mimo to detekcji dokonuje drugi czujnik, jednak oba czujniki w takich przypadkach działają.

3. W rozprawie nie określa się, na ile złe warunki atmosferyczne (deszcz, śnieg, mgła) mogą wpływać na działanie proponowanego podejścia. Zestaw NuScenes wydaje się do pewnego stopnia zawierać obrazy rejestrowane przy niekorzystnych warunkach pogodowych. Zestaw danych KITTI dotyczy wprawdzie jedynie dobrych warunków pogodowych, jednak np. w pracy M. J. Mirza *et al.*, "Robustness of Object Detectors in Degrading Weather Conditions," *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, USA, 2021, pp. 2719-2724 raportuje się eksperymenty symulujące złe warunki pogodowe poprzez odpowiednią modyfikację obrazów.

4. W rozdz. 6.2.2 podano, że funkcja strat używana podczas uczenia sieci jest kombinacją członów odpowiadających klasyfikacji i regresji, nie specyfikując jednak, czy chodzi tu o zwykłą sumę (wtedy problemem może być różna wielkość obu członów), czy też o sumę ważoną (wtedy należałoby określić, jak wybrano wagi członów tak, aby jeden z nich nie zdominował zagregowanej funkcji strat).

5. W rozprawie skupiono się na detekcji i klasyfikacji obiektów, jednak stanowią one tylko pewną część układu sterowania i systemu podejmowania decyzji. W literaturze spotyka się fuzję danych z kamery i radaru na potrzeby predykcji trajektorii (A. Benterki, M. Boukhnifer, V. Judalet and C. Maaoui, "Artificial Intelligence for Vehicle Behavior Anticipation: Hybrid Approach Based on Maneuver Classification and Trajectory Prediction," in *IEEE Access*, vol. 8, pp. 56992-57002, 2020) lub nawigacji pojazdów (P. Cai, S. Wang, Y. Sun and M. Liu, "Probabilistic End-to-End Vehicle Navigation in Complex Dynamic Environments With Multimodal Sensor Fusion," in *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4218-4224, July 2020). Czy podejście zaproponowane w rozprawie jest na tyle perspektywiczne, że można byłoby je przenieść również na tego typu zastosowania?

6. Dywagacje rozdz. 5.2 dotyczące transformacji układu współrzędnych kamery do układu współrzędnych pojazdu wydają się niepotrzebnie skomplikowane. Sprowadzają się one do złożenia dwóch prostych obrotów (obrót o 180 stopni wokół osi z oraz następującym po nim obrót o 90 stopni wokół bieżącej osi y), co sam Autor stwierdza na str. 51, przy czym należałoby jeszcze uwzględnić przesunięcie początków układów współrzędnych. Wynikałoby z tego, że jeśli jakiś punkt ma w nowym układzie współrzędne (x^1, y^1, z^1), to w starym układzie będzie miał współrzędne $(a - z^1, -y^1, b - x^1)$, gdzie: a, b – przesunięcia. Trochę niezrozumiałym jest potrzeba odwoływania się aż do pojęcia kwaternionów, a w świetle poprzedniego zdania kategoryczne stwierdzenie „it is worth emphasizing that achieving this level of alignment is not possible with arbitrary combinations of permutations or transpositions of the input tensor dimensions" (u dołu str. 51) nie wydaje się uzasadnione. Podobnie, trochę zaskakujące jest stwierdzenie „in some cases, the rotation may result in negative indexes" (u dołu str. 50). Prawdopodobnie wynika ono z ograniczenia się tylko do obrotów i pominięcia przesunięć a, b.

7. Pewien niedosyt pozostawia przyjęty sposób opisu proponowanych rozwiązań, redukujący do minimum używanie wzorów matematycznych lub schematów algorytmów. W to miejsce pojawia się opis słowny, siłą rzeczy nie zawsze precyzyjny. Dotyczy to również sposobów agregacji opisywanych w rozdz. 5. Cechą raportowanych badań naukowych powinna być precyzja i replikowalność. Oczywiście, przy tak złożonych systemach, jak opisywane sieci głębokie, całkowite odtworzenie sposobu prowadzenia badań na podstawie ich opisu jest

praktycznie niemożliwe, jednak opis słowny, czasem lakoniczny lub niejednoznaczny, takie niebezpieczeństwo mocno potęguje. Precyzyjniejszy opis formalny przydałby się zwłaszcza w opisach oryginalnych pomysłów Autora (rozdz. 5.4.3 i rozdz. 9.1), kiedy brakuje pozycji innych literaturowych mogących naświetlić więcej szczegółów.

8. Opisując wkład rozprawy do obecnego stanu wiedzy (str. 3) Autor umieścił dokonanie obszernego przeglądu literaturowego, co nie jest udanym pomysłem. Dokonanie przeglądu istniejącej literatury jest obowiązkiem każdego autora rozprawy doktorskiej lub monografii naukowej i nie kwalifikuje się tego jako szczególnego osiągnięcia.

9. Przy odwołaniach literaturowych stosuje się styl Harvard/Chicago, podając nazwisko głównego autora i rok wydania. Niestety, lista literatury na str. 127 jest numerowana i nie posortowana alfabetycznie według nazwisk, co mocno utrudnia odnalezienie konkretnej publikacji. Ponadto lista autorów publikacji wieloautorskich jest najczęściej niepełna (w miejsce brakujących autorów stosuje się skrót „et al."). Ta uwaga krytyczna dotyczy to również listy publikacji samego Autora przedstawionej na str. 4, co utrudnia oszacowanie jego wkładu w przygotowanie danej publikacji.

Przyznaję jednak, że powyższe uwagi i komentarze nie mają jednak przesadnego wpływu na ogólną opinię o recenzowanej dysertacji, którą zdecydowanie oceniam jako bardzo wartościową.

V. Podsumowanie

Uwzględniając wyżej wymienione uwagi i komentarze oraz całość rozprawy doktorskiej wraz z oryginalnymi osiągnięciami naukowo-badawczymi stwierdzam, że

1. **recenzowana rozprawa doktorska Pana mgr inż. Daniela Dworaka spełnia wszystkie wymagania Ustawy z dnia 20 lipca 2018 r.** ***Prawo o szkolnictwie wyższym i nauce*** **(Dz. U. 2023, poz. 742) w odniesieniu do rozpraw doktorskich;**

2. **w związku z tym wnoszę o dopuszczenie Autora rozprawy do dalszych, przewidzianych przepisami, etapów przewodu doktorskiego.**

Dariusz Michalski