Journal of Electronical & Cyber Defence
Vol. 9, No. 4, 2022, Serial No. 36

The Semantic Segmentation of Autonomous Vehicles Images with the Teacher-Student Technique

Amir Khosravian[1], Masoud Masih-Tehrani[2], Abdollah Amirkhani[3*]

*Assistant Professor, School of Automotive Engineering, Iran University of Science and Technology

(Received: 05/10/2021, Accepted: 10/07/2021)

ABSTRACT

Semantic segmentation is one of the most common image processing outputs for vision-based autonomous vehicles. Deep neural networks require large-scale data in order to learn new environment features with diverse domains. While the approach of a great deal of papers is based on supervised learning, in this paper, semantic segmentation has been implemented by taking advantage of the semi-supervised learning method. To be more specific, in this study the teacher-student technique is utilized to establish a connection for the interaction between the deep learning models. First, the DABNet and ContextNet models are trained as our teacher networks with the BDD100K database. Regarding the significance of generalization and robustness of models in autonomous vehicles, these criteria of the teacher models have been evaluated by simulations in CARLA software. Finally, teacher networks train the FastSCNN model automatically using the Cityscapes database without any human interference. In contrast with other semi-supervised approaches, the existence of two different databases with noticeable amount of domain-shift effect would challenge the student-teacher technique even more. The results indicate that student's performance in classes such as vehicles, pedestrians, and road, which are the highest priority classes to detect, has only 1.2%, 3%, and 3.8% accuracy difference, respectively. Also, there is a 4.5% drop for the model's mean intersection over union accuracy between the teacher's performance and a similar model trained with an entirely supervised method. Also, the mean accuracy for the student model has only 4.5% difference in performance with a model whose data base needs a long time for preparation.

Keywords: Autonomous vehicles, Convolutional neural networks, Semantic segmentation, Teacher-student technique

*Corresponding Author Email: amirkhani@iust.ac.ir

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۴، زمستان ۱۴۰۰، ص ۱۹- ۱

علمی- پژوهشی

قطعه‌بندی معنایی تصاویر خودروهای خودران با بهره‌گیری از روش معلم- دانش‌آموز

امیر خسرویان[۱]، مسعود مسیح طهرانی[۲]، عبدالله امیرخانی[۳*]

۱- کارشناسی ارشد، ۲ و ۳- استادیار، دانشکده مهندسی خودرو، دانشگاه علم و صنعت، تهران، ایران

(دریافت: ۱۳۹۹/۰۷/۱۴، پذیرش: ۱۴۰۰/۰۴/۱۹)

چکیده

قطعه‌بندی معنایی یکی از رایج‌ترین خروجی‌های پردازش تصویری برای خودروهای خودران مجهز به بینایی است. مدل‌های مبتنی بر یادگیری عمیق جهت یاد گرفتن ویژگی‌های محیطی جدید و با دامنه متفاوت نیازمند در اختیار داشتن انبوهی از داده هستند. اما فرآیند برچسب‌گذاری دستی این حجم از داده توسط انسان بسیار زمان‌بر خواهد بود. در حالی که رویکرد بسیاری از مقالات مبتنی بر آموزش مدل‌های یادگیری عمیق با روش نظارتی است، در این مقاله از روش نیمه نظارتی جهت اعمال قطعه‌بندی معنایی بهره گرفته می‌شود. به‌طور دقیق‌تر در این پژوهش، روش معلم- دانش‌آموز جهت برقراری تعامل میان مدل‌های یادگیری عمیق به کار گرفته می‌شود. در ابتدا مدل‌های DABNet و ContextNet در جایگاه معلم با استفاده از پایگاه داده BDD100K آموزش داده می‌شوند. با توجه به اهمیت قابلیت تعمیم پذیری و مقاوم بودن مدل‌های مورد استفاده در خودروهای خودران، این معیارهای شبکه‌های معلم با شبیه‌سازی در نرم‌افزار CARLA مورد ارزیابی قرار گرفته‌اند. سپس شبکه‌های معلم، پایگاه داده Cityscapes را به‌طور کامل و بدون دخالت انسان در فرآیند آموزش با بهره‌گیری از یادگیری نیمه- نظارتی به مدل FastSCNN آموزش داده‌اند. برخلاف سایر رویکردهای نیمه- نظارتی، وجود دو پایگاه داده با اختلاف دامنه قابل توجه، روش معلم- دانش‌آموز را بیشتر به چالش خواهد کشید. نتایج نشان می‌دهد عملکرد مدل دانش‌آموز در کلاس‌هایی نظیر خودرو، انسان و جاده که شناسایی آن‌ها از مهم‌ترین اولویت‌های خودرو خودران است به‌ترتیب به میزان ۱/۲٪، ۳٪ و ۳/۸٪ با برچسب‌گذاری دستی اختلاف دارد. همچنین میانگین دقت مدل دانش‌آموز نیز تنها ۴/۵٪ اختلاف عملکرد با مدلی دارد که آماده‌سازی پایگاه داده آن نیازمند صرف زمان بسیار زیاد است.

کلید واژه‌ها: خودرو خودران، شبکه‌های عصبی پیچشی، قطعه‌بندی معنایی، روش معلم- دانش‌آموز

۱- مقدمه[۱]

مطابق با گزارش سازمان ملی مدیریت ایمنی ترافیک بزرگراه، خطاهای انسانی ۹۴٪ از تصادفات دنیا را سبب می‌شود [۱]. به‌همین علت است که بسیاری از محققان در سراسر جهان در تلاش هستند تا با ارائه راهکارهای متفاوت، عامل انسان را از خودرو حذف نموده و آینده خودرو را به سمت خودروهای خودران ببرند [۴- ۲]. با حذف عامل راننده، خودرو نیازمند حسگرهایی جهت ادراک محیط پیرامون خود خواهد بود. این حسگرها موظفند مسیر حرکت و موانع موجود در مسیر را شناسایی کرده و در اختیار کنترل‌کننده خودرو قرار دهند [۵]. بینایی از جمله حسگرهای کاربردی و رایج جهت استفاده در خودروهای خودران است [۶]. علت این امر را می‌توان در هزینه کم و دقت بالای طبقه‌بندی این حسگر دانست. تصاویر دریافتی حسگر بینایی توسط الگوریتم‌های متفاوت یادگیری ماشین (به خصوص مبتنی بر یادگیری عمیق) مورد پردازش قرار گرفته و مسیر و موانع پیرامون خودرو مشخص می‌گردد. قطعه‌بندی معنایی به‌عنوان یکی از مهم‌ترین خروجی‌های بینایی ماشین است که قادر است مسیر و موانع موجود در تصویر را در مقیاس پیکسلی برای خودرو خودران مشخص سازد. امروزه با پیشرفت یادگیری عمیق، شبکه‌های عصبی پیچشی[۲۱](CNN) قادر هستند تصاویر دریافتی را با دقت و سرعت بالا پردازش نموده و خروجی را در اختیار کنترل‌کننده خودرو قرار دهند [۷]. علت اصلی شکوفایی روش‌های مبتنی بر یادگیری عمیق در سال‌های اخیر را می‌توان در وجود انبوه داده جهت آموزش جامع یک مدل دانست. اگرچه با حذف راننده در خودروهای خودران، یکی از عامل‌های اصلی تصادف‌های حذف می گردد، اما چالش دیگری وجود دارد. در واقع استفاده از الگوریتم‌های یادگیری عمیق برای پردازش اطلاعات، خودروهای خودران را در معرض خطر حمله‌های سایبری قرار می‌دهد [۸]. از سوی دیگر با گسترش تکنولوژی خودروهای متصل، ارتباط خودرو با سایر خودروها (V2V) و همچنین محیط اطراف (V2I) نیز افزایش خواهد یافت که به‌طور مشابه در این ارتباط‌ها نیز امکان وجود حملات سایبری بالا خواهد بود [۹]. مرجع [۸] حملات سایبری در خودروهای

*رایانامه نویسنده مسئول: amirkhani@iust.ac.ir

[1] Convolutional Neural Network

ناشر: دانشگاه جامع امام حسین (ع)

خودران را در سه بخش سیستم کنترل خودران، اجزاء سیستم رانندگی و ارتباط خودرو با همه‌چیز (V2X) مورد بررسی قرار داده و روش‌های به‌کار گرفته شده جهت دفاع در برابر این حملات را مرور کرده است. این موضوع نشان دهنده نقش پررنگ امنیت اطلاعات در نوع معماری انتخابی، الگوریتم آموزش شبکه، برچسب‌های نسبت داده شده و تشخیص‌های بلادرنگ خروجی شبکه در خودروهای خودران مبتنی بر هوش مصنوعی و یادگیری عمیق است.

نکته قابل توجه دیگر در خودروهای خودران آن است که CNNها هنگام رویارویی با تصاویری خارج از دامنه آموزش خود همچنان با افت عملکرد چشم‌گیری مواجه می‌شوند [۱۲– ۱۰]. بنابراین نیاز است تا CNNها بخش قابل توجهی از دامنه محیطی که انتظار می‌رود در آن مورد ارزیابی قرار بگیرند را در طی فرآیند آموزش به خوبی مشاهده نمایند [۱۳]. در غیر این صورت، مدل‌های یادگیری عمیق با پدیده اختلاف دامنه میان تصاویر آموزش و آزمایش خود مواجه شده که در نهایت منجر به افت دقت قابل توجهی در عملکرد آن‌ها خواهد گشت. با این وجود در اکثر پژوهش‌ها، محققین همچنان از پایگاه داده‌های رایج معرفی شده نظیر Cityscapes [۱۴]، BDD100K [۱۵] و KITTI [۱۶] استفاده می‌نمایند. علت این امر آن است که آماده ساختن یک پایگاه داده اختصاصی و جدید نیازمند صرف زمان بسیار زیاد در راستای انجام برچسب‌گذاری یکایک پیکسل‌ها در تمام تصاویر پایگاه داده آموزش و ارزیابی است.

برقراری ارتباط آموزشی میان مدل‌های مبتنی بر یادگیری عمیق با استفاده از روش معلم- دانش‌آموز می‌تواند پاسخی در راستای آموزش یک پایگاه داده جدید به مدل بدون صرف زمان بسیار زیاد جهت برچسب‌گذاری تصاویر آن باشد [۱۷]. در این رویکرد یک مدل یادگیری عمیق که بر پایگاه‌های داده رایج آموزش داده شده است (شبکه معلم) بر تصاویر پایگاه داده جدید قطعه‌بندی معنایی را اعمال می‌نماید. مدل یادگیری عمیق دوم (دانش‌آموز) بر تصاویر قطعه‌بندی شده توسط معلم آموزش دیده و پایگاه داده جدید را توسط پیش‌بینی‌های شبکه معلم فرا می‌گیرد. باید توجه داشت که ذات دو شبکه معلم و دانش‌آموز با یکدیگر در حیطه خودروهای خودران تفاوت اساسی دارد. شبکه معلم شبکه‌ای عمیق است که تنها قرار است دامنه جدید را به شبکه دانش‌آموز آموزش دهد. این شبکه می‌تواند از استخوان‌بندی‌های عمیق بهره ببرد. شبکه‌های DeeplabV3 [۱۸] و همچنین DeeplabV3+ [۱۹] با استخوان‌بندی‌های عمیقی نظیر ResNet [۲۰] می‌توانند نمونه شبکه‌های مناسبی برای استفاده در جایگاه معلم باشند [۲۱]. عمق، تعداد پارامتر و همچنین درجه آزادی بالا به این خانواده از شبکه‌ها اجازه می‌دهد تا با دقت قابل قبولی اقدام به آموزش شبکه دانش‌آموز بنماید. همچنین باید توجه داشت که لزومی نیست شبکه معلم قادر باشد پردازش بی‌درنگ انجام دهد اما قابلیت تعمیم و مقاوم بودن این شبکه هنگام آموزش آن باید به نحوی حفظ شده باشد که هنگام رویارویی با پایگاه داده جدید با افت عملکرد شدید مواجه نشود. در نقطه مقابل، شبکه دانش‌آموز نمی‌تواند از لایه‌های بسیار عمیق استفاده نماید چرا که این شبکه باید قادر باشد پردازش بی‌درنگ انجام دهد. این شبکه بر خودرو در حال حرکت استفاده می‌گردد و در صورتی که نتواند سرعت پردازشی مورد نیاز یک خودرو خودران (بیش از ۴۰ فریم بر ثانیه) را پوشش دهد، ممکن است خسارات جبران ناپذیر مالی و جانی به وجود آورد [۲۲]. با توجه به این که خودروها به‌طور مستقیم با جان انسان‌ها در ارتباط هستند، CNNهای استفاده شده در آن‌ها باید به‌طور همزمان دارای دقت و سرعت پردازشی بالا باشند [۲۳].

سایر بخش‌های این مقاله به شرح زیر مطرح می‌گردند. در بخش ۲ برخی منابع مختصر در حوزه روش معلم دانش‌آموز و کاربرد آن در خودروهای خودران مرور شده است. در بخش ۳ پایگاه‌های داده و مدل‌های مبتنی بر یادگیری عمیق انتخابی به همراه پارامترهای آموزشی مورد بحث واقع شده‌اند. در بخش ۴ نیز شبکه‌های معلم و دانش‌آموز ارزیابی شده و نتایج عملکرد آن‌ها تفسیر می‌شوند. در نهایت، جمع‌بندی نتایج مقاله در بخش ۵ انجام شده است.

۲- مروری بر منابع

پژوهش‌های قابل توجهی در راستای تکامل ایده معلم- دانش‌آموز به منظور شناسایی دامنه تصاویر جدید صورت گرفته است. ژو و همکاران [۲۴] در سال ۲۰۲۰ با روش آموزش خود بهبود قابل توجهی در روش آموزشی معلم- دانش‌آموز صورت دادند. ارزیابی بر پایگاه‌های داده Cityscapes، KITTI [۲۵] و Camvid [۲۶] نشان می‌دهد روش مبتنی بر نظارت محققین قادر است با دقت مدل‌های به‌روز به‌طور کامل رقابت کند. ایده معلم- دانش‌آموز توسط چن و همکاران [۲۷] در سیستم نیمه- نظارتی مورد استفاده قرار گرفته است. روش آموزش با استفاده از داده‌ها با برچسب‌گذاری انسان و همچنین معلم در چندین گام توسط محققین تکرار شده است. ارزیابی توسط پایگاه داده Cityscapes در حیطه‌های متفاوت پردازش تصویر نظیر قطعه‌بندی معنایی، نمونه‌ای و پانوپتیک نشان داده است که استفاده از روش معلم- دانش‌آموز توانسته است مدل را به دقتی با قابلیت استناد بالا برساند. در سال ۲۰۲۰، زی و همکاران [۲۸] نیز به‌طور مشابه از روش معلم- دانش‌آموز در راستای چرخه گام‌به‌گام میان شبکه‌های معلم و دانش‌آموز استفاده کردند. در

ابتدا شبکه معلم بر داده‌هایی با برچسب‌گذاری دستی آموزش داده می‌شود. سپس توسط این شبکه، بر توده عظیمی از داده متشکل از ۳۰۰ میلیون تصویر برچسب‌گذاری خودکار اعمال شده که برای آموزش دانش‌آموز مورد استفاده قرار می‌گیرند. همچنین جهت افزایش قابلیت تعمیم شبکه دانش‌آموز، از روش‌های متفاوت تولید داده و همچنین اعمال نویزهای گوناگون بر تصاویر آموزشی استفاده گردیده است. نویزهای مذکور باعث می‌شوند تا مدل CNN به‌جای استخراج و یادگیری ویژگی‌های مبتنی بر بافت تصویر به یادگیری شکل اشیاء موجود در تصویر بپردازد که در نهایت منجر به افزایش قابلیت تعمیم آن می‌گردد. در مرحله بعد شبکه دانش‌آموز آموزش داده شده خود در جایگاه معلمی جدید قرار گرفته و حلقه آموزشی از ابتدا تکرار می‌گردد. استفاده از تصاویر مصنوعی پایگاه‌های داده‌ای نظیر SYNTHIA [۲۹] از دیگر رویکردهای مورد نظر محققین بوده است. اما باید توجه داشت که این تصاویر به شدت با پدیده اختلاف دامنه میان تصاویر مصنوعی در مقایسه با دنیای واقعی مواجه هستند. پژوهش‌هایی نظیر مراجع [۳۳- ۳۰] تلاش کرده‌اند تا با ادغام مستقیم تصاویر مصنوعی و واقعی در پایگاه داده آموزش به این چالش پاسخ دهند. با این وجود باید توجه داشت که اختلاف دامنه میان تصاویر پایگاه داده آموزش خود می‌تواند منجر به سردرگمی مدل در استخراج ویژگی گردد. در نقطه مقابل پن و همکاران [۳۴] در سال ۲۰۲۰ با رویکردی خود- نظارت تلاش کردند انطباق دامنه مدل را هنگام گذر از تصاویر مصنوعی به واقعی بهبود بخشند. از دیگر رویکردهایی که با محوریت استفاده از تصاویر مصنوعی و انطباق دامنه آن با استفاده از تصاویر واقعی اقدام به بهره‌برداری از روش معلم- دانش‌آموز داشته‌اند می‌توان به میشیلی و همکاران [۳۵] ارجاع داد. پایگاه‌های داده SYNTHIA و GTA5 [۳۶] برای تصاویر مصنوعی و پایگاه‌های داده Cityscapes و Mapillary [۳۷] به‌عنوان منبع تصاویر واقعی در این مرجع مورد استفاده قرار گرفته‌اند.

در اغلب مقالاتی که مرور گردید، رویکرد نیمه- نظارتی در پایگاه داده ثابت انجام گرفته است و تصاویری که شبکه معلم برچسب‌گذاری می‌نماید به‌طور عمده متعلق به همان دامنه تصاویری هستند که شبکه معلم از آن‌ها آموزش دیده است. به همین علت قابلیت تعمیم این رویکردها برای هنگامی که شبکه معلم با دامنه تصاویری کاملا متفاوت مواجه می‌شود به‌طور کامل بررسی نشده است. به‌طور دقیق‌تر مشخص نیست که آیا ایده معلم- دانش‌آموز هنگام مواجه با پدیده اختلاف دامنه شدید میان تصویر فاقد برچسب‌گذاری و تصاویر آموزش خود، همچنان قابل استناد است؟ در این مقاله با استفاده از رویکرد معلم- دانش‌آموز و آموزش خودکار اقدام به پیاده‌سازی قطعه‌بندی معنایی می‌شود. جهت پاسخگویی به چالش مطرح شده در رابطه با اختلاف دامنه، ابتدا شبکه‌های معلم با استفاده از پایگاه داده BDD100K آموزش داده شده و از نقطه نظر دقت و سرعت پردازشی با یکدیگر مقایسه می‌شوند. سپس قابلیت تعمیم و مقاوم بودن این شبکه‌ها مورد بررسی قرار می‌گیرد. در نهایت پایگاه داده Cityscapes توسط مدل‌های معلم بدون دخالت انسان به مدل دانش‌آموز آموزش داده شده و نتایج آن با هنگامی که این آموزش توسط انسان صورت می‌گیرد مقایسه می‌گردد. با توجه به اختلاف دامنه قابل توجه میان دو پایگاه داده ذکر شده، این مقاله بهره بردن از انبوه داده‌ها را در روش معلم- دانش‌آموز به مراتب بهتر از پژوهش‌هایی که تنها از یک پایگاه داده استفاده کرده‌اند نشان می‌دهد.

۳- پایگاه داده و مدل‌های یادگیری عمیق

تعداد قابل توجهی از پژوهش‌ها از رویکرد نظارت شده جهت آموزش مدل استفاده می‌نمایند [۳۹ و ۳۸]. به همین علت علاوه بر افزایش زمان آماده ساختن پایگاه داده، امکان استفاده از انبوه داده‌های فاقد برچسب‌گذاری را نیز نخواهند داشت. در این مقاله جهت برطرف کردن این محدودیت، از چندین پایگاه داده و مدل‌های مختلف یادگیری عمیق برای پیاده‌سازی روش معلم- دانش‌آموز استفاده شده است. در این قسمت پایگاه‌های داده مورد استفاده و دلایل استفاده از آن‌ها شرح داده خواهد شد. همچنین مدل‌های مبتنی بر یادگیری عمیق که به ترتیب برای جایگاه‌های معلم و دانش‌آموز انتخاب شده‌اند معرفی شده و معماری آن‌ها شرح داده خواهد شد. در انتها نیز پارامترهای مورد استفاده و کلاس‌های انتخابی جهت آموزش به شبکه و اعمال قطعه‌بندی معنایی مورد بحث قرار می‌گیرند.

۳-۱- پایگاه داده

دو پایگاه داده با اختلاف دامنه قابل توجه نسبت به یکدیگر جهت استفاده از روش معلم- دانش‌آموز مورد نیاز است. یکی از این پایگاه‌های داده مسئول مستقیم آموزش شبکه معلم با استفاده از برچسب‌گذاری دستی است. پایگاه داده دیگر که فرض می‌گردد برچسب‌گذاری دستی آن موجود نیست، توسط شبکه معلم برچسب‌گذاری خودکار دریافت کرده و به شبکه دانش‌آموز آموزش داده می‌شود. شایان ذکر است که وجود برچسب‌گذاری دستی به‌عنوان حقیقت محض تصاویر در پایگاه داده دانش‌آموز نیز الزامی است. علت این امر آن است که شبکه CNN دانش‌آموز به‌طور مجزا توسط برچسب‌گذاری دستی نیز آموزش داده شود تا بتوان مقایسه منصفانه‌ای از عملکرد مدل دانش‌آموز داشت.

شکل (۱): نمونه تصاویری از پایگاه داده Cityscapes (سطر اول) [۱۴] و BDD100K (سطر دوم) [۱۵].

با توجه به مطالب توضیح داده شده، BDD100K به‌عنوان پایگاه داده شبکه معلم انتخاب گشته است. این پایگاه داده تشکیل شده از ۱۰۰۰۰۰ ویدئو جمع‌آوری شده در شرایط بسیار متنوع رانندگی است. BDD100K به‌عنوان یکی از بزرگترین پایگاه‌های داده موجود در حیطه پردازش تصویر، پیچیدگی‌های زیادی را در خود جای داده است. دلایل متعددی ما را به استفاده از این پایگاه داده جهت انجام قطعه‌بندی معنایی متمایل کرده است. هر چه قابلیت تعمیم پایگاه داده در ارائه تصاویر بیشتر باشد بدیهی است که مدل یادگیری عمیق می‌تواند بهتر آموزش دیده و عملکرد دقیق‌تری از خود برجای بگذارد. ویژگی‌های یک پایگاه داده عمومی را می‌توان در ارائه داده در شهرها، آب و هواها، تنظیم دوربین و صحنه‌های تصویربرداری متفاوت خلاصه کرد. پایگاه داده‌ای نظیر KITTI به علت عدم پشتیبانی از شهرها و وضعیت‌های آب و هوایی مختلف انتخاب نگشته است. این در حالی است که BDD100K در چندین شهر مختلف نظیر نیویورک، برکلی و سانفرانسیسکو جمع‌آوری شده است. همچنین BDD100K در پایگاه داده مربوط به قطعه‌بندی معنایی خود چندین شرایط آب و هوایی متنوع را مانند روز، شب، باران و برف پوشش داده است. بخش استفاده شده در مقاله این پایگاه داده که به قطعه‌بندی معنایی اختصاص یافته است شامل ۷۰۰۰ تصویر جهت آموزش و ۱۰۰۰ تصویر جهت آزمایش است.

در نقطه مقابل، پایگاه داده Cityscapes برای آموزش دانش‌آموز به کار گرفته شده است. این پایگاه داده غالباً در کشور آلمان جمع‌آوری شده است. پایگاه Cityscapes نیز همانند BDD100K در چندین شهر مختلف تصویربرداری شده است، با این تفاوت که مدل‌های متنوع آب و هوایی را پشتیبانی نمی‌کند. نخستین علت انتخاب این پایگاه داده آن است که تصاویر Cityscapes از آنجا که در آلمان جمع‌آوری شده‌اند اختلاف دامنه قابل توجهی با BDD100K (جمع‌آوری شده در آمریکا) دارد. دلیل مهم دیگر در این انتخاب آن است که Cityscapes به علت قدیمی‌تر بودن (۲۰۱۶) در مقایسه با BDD100K توسط مراجع و محققین بسیاری در سراسر جهان مورد استفاده قرار گرفته و بسیاری از شبکه‌های به روز و مطرح با استفاده از این پایگاه داده ارزیابی و صحت‌سنجی شده‌اند. بخش نخست این پایگاه داده ۲۹۷۵ تصویر برای آموزش، ۵۰۰ تصویر جهت صحت‌سنجی و ۱۵۲۵ تصویر جهت آزمایش با حجمی در حدود ۱۲ گیگابایت دارد. همچنین فایل تکمیل‌کننده آموزش مدل این پایگاه نیز ۲۰۰۰۰ تصویر را با برچسب‌گذاری تقریبی پوشش داده که بیش از ۴۰ گیگابایت حجم دارد. در این مقاله از هر دو بخش داده‌های موجود به‌طور کامل استفاده شده است. شکل (۱) نمونه تصاویری از پایگاه‌های داده BDD100K و Cityscapes را نمایش می‌دهد.

۳-۲- انتخاب مدل

در این بخش به توصیف مدل‌های انتخابی در جایگاه معلم و دانش‌آموز پرداخته می‌شود. لازم به یادآوری است که استفاده از مدلی که قابلیت پردازش بی‌درنگ را نداشته باشد برای جایگاه دانش‌آموز و مدلی که قابلیت تعمیم خوبی را دارا نباشد در جایگاه معلم توصیه نمی‌شوند. بنابراین، DABNet [۴۰] و ContextNet [۴۱] به‌عنوان معلم و FastSCNN [۴۲] به‌عنوان شبکه دانش‌آموز در این مقاله مورد استفاده قرار گرفته‌اند.

در سال ۲۰۱۹، DABNet برای برقرار کردن تناسب بین دقت و سرعت پردازش معرفی شد. در مقاله مرجع، محققین با ابداع ماژول DAB اقدام به ساخت شبکه‌ای گلویی با لایه‌های عمیق نامتقارن کردند. این شبکه تنها به اندازه مورد نیاز زمینه پردازشی تولید کرده و از اطلاعات متنی استخراج شده از ویژگی‌های تصویر به صورت کاملا متراکم استفاده می‌نماید. به علت تعداد کم پارامترهای تشکیل‌دهنده این شبکه، امکان آموزش کامل وجود دارد و احتیاجی به استفاده از مدل‌های پیش- آموزش دیده نیست. طبق گزارش‌های موجود در مرجع [۴۰]، دقت و سرعت پردازش این شبکه به ترتیب برابر با ٪۷۰/۱ (در معیار mIoU) و ۱۰۴ فریم بر ثانیه است. همچنین تعداد

پارامترهای شبکه مذکور نیز برابر با ۰/۷۶ میلیون است. ماژول DAB ساختار مشابهی با ResNet [۲۰] را دنبال می‌کند. هر ساختار گلویی ابتدا تعداد کانال‌های ویژگی را به نصف تقسیم نموده و کانال‌های اصلی را با پیچشی مبتنی بر نقاط ترمیم و بازگردانی می‌کند. در ساختار گلویی ماژول DAB، ابتدا پیچشی با فیلتر یا کرنل ۳ × ۳ از هر ماژول عبور می‌کند. نکته قابل توجه آن است که کرنل پیچشی ۱ × ۱ تعداد پارامترهای کمتری را دارا است، بنابراین منجر به ساختن مدلی عمیق خواهد شد. بدیهی است که عمیق بودن شبکه باعث افزایش زمینه پردازشی شده و توانایی شبکه در استخراج ویژگی‌های پیچیده‌تر را بهبود می‌بخشد. اما چالش اصلی عمیق کردن شبکه آن است که با وسیع شدن زمینه پردازشی، سرعت پردازش پایین آمده و شبکه نیازمند سخت‌افزار قدرتمندتری خواهد بود. بنابراین به علت اجتناب از ساختن مدلی بسیار عمیق و نیز بالا بردن زمان پردازشی، در گلویی DAB از کرنل ۳ × ۳ استفاده گشته است. در ماژول DAB پس از اولین پیچش، تعداد کانال‌ها تنها به نصف تقسیم شده است چرا که بیشینه کانال موجود در DABNet برابر با ۱۲۸ بوده و با هزاران کانال موجود در ResNet [۲۰] قابل مقایسه نیست. بنابراین جهت حفظ اطلاعات فضایی، کانال‌ها بیش از حد فشرده نشده‌اند.

معماری FastSCNN نیز از دیگر شبکه‌های جدید مبتنی بر پردازش بی‌درنگ است. ماژول اصلی و منحصر به فرد این شبکه، یادگیری در کاهش نمونه نام داشته که ویژگی‌های سطح پایین را از تصاویری با چندین اندازه مختلف به‌طور همزمان استخراج می‌نماید. این شبکه، اطلاعات فضایی را که از تصاویر با اندازه بزرگ استخراج شده‌اند با ویژگی‌های عمیق استخراج شده از تصاویر با اندازه کوچک ترکیب نموده که منجر به دقت mIoU معادل با ۶۸٪ و سرعت ۱۲۳/۵ فریم بر ثانیه شده است. معماری FastSCNN از ساختار دو انشعابی [۴۴ و ۴۳] و همچنین شبکه‌های مبتنی بر رمزگذار-رمزگشا [۴۶ و ۴۵] الهام گرفته است. این معماری از ۴ بخش اصلی تحت عناوین یادگیری به کاهش نمونه، استخراج‌کننده ویژگی جهانی، ترکیب ویژگی و طبقه‌بند تشکیل شده است. در مرحله نخست، تنها از سه لایه جهت اطمینان استخراج ویژگی‌های سطح پایین استفاده شده است. لایه اول از نوع پیچش معمولی و دو لایه بعدی از نوع پیچش عمیق تفکیک‌پذیر هستند. در هر سه لایه نامبرده از انتقال کرنل ۲ به‌همراه نرمال‌سازی دسته ورودی و در نهایت فعال‌ساز Relu استفاده گشته است. همچنین اندازه کرنل‌های پیچشی نیز در هر سه لایه ۳×۳ انتخاب شده است. استخراج‌کننده ویژگی جهانی جهت استخراج ویژگی‌های کلی موجود در تصاویر مورد استفاده قرار می‌گیرد. در تضاد با روش‌های دو-انشعابی رایج که با استفاده از تصاویری با اندازه کوچک اعمال می‌گردند، معماری FastSCNN خروجی ماژول اول را دریافت می‌نماید. این خروجی تنها یک هشتم اندازه تصویر ورودی است. این شبکه نیز از بلوک گلویی [۴۹] که از پیچش عمیق تفکیک‌پذیر موثری با توجه به قرارگرفتن در حیطه پردازش بی‌درنگ و کاهش تعداد پارامترهای شبکه استفاده می‌نماید. در نهایت با استفاده از ماژول استخراج هرمی اطلاعات محلی موجود در مناطق مختلف تصویر با یکدیگر جمع می‌شوند. در ماژول ترکیب ویژگی به جمع ساده ویژگی‌ها با یکدیگر جهت بهینه بودن زمان پردازشی مانند [۵۰] اکتفا شده است. در ماژول نهایی (طبقه‌بند) از دو لایه پیچشی عمیق تفکیک‌پذیر و نیز یک لایه پیچشی معمولی استفاده می‌گردد. درنهایت به علت استفاده از بهینه‌ساز گرادیان نزولی تصادفی[1] (SGD)، از طبقه‌بند softmax جهت انجام پیش‌بینی شبکه استفاده می‌گردد.

جهت بررسی عمیق‌تر تاثیر شبکه معلم در روش معلم- دانش‌آموز از شبکه دیگری نیز تحت عنوان ContextNet استفاده گردیده است. این معماری با محوریت فشرده‌سازی شبکه عصبی و بازنمایی هرمی بنا شده است. دقت و سرعت پردازشی مدل مذکور بر پایگاه داده Cityscapes به ترتیب برابر با ۶۶/۱٪ و ۴۱/۹ فریم بر ثانیه بوده است. اندازه تمامی تصاویر ۱۰۲۴ × ۲۰۴۸ و ارزیابی مدل توسط Nvidia Titan X Maxwell انجام شده است. این شبکه نیز همانند معماری پیشین، از پیچش عمیق تفکیک‌پذیر به همراه پیچش معمولی با اندازه کرنل ۱ × ۱ جهت کاهش تعداد پارامترهای شبکه استفاده می‌کند. باید توجه داشت که این شبکه قطعه‌بندی بهینه از تصاویر با اندازه کوچک ارائه کرده که در انتها این نتایج با ریز شبکه دیگری که بر تصاویر بزرگ پردازش انجام می‌دهد ترکیب شده تا در نهایت جزئیات قطعه‌بندی محیط نمایش داده شود. ریز شبکه ContextNet نیز مطابق با FastSCNN از بلوک‌های ساختار گلویی به همراه پیچش عمیق تفکیک‌پذیر بهره می‌برد. در معماری ContextNet به‌طور کلی از ۳۸ لایه موثر جهت توصیف فضای ویژگی‌ها استفاده می‌گردد. در نقطه مقابل ریز شبکه اعمالی بر تصاویر بزرگ تا حد ممکن از نوع کم عمق است چرا که استفاده از ریز شبکه‌ای عمیق برای تصاویر با اندازه بزرگ می‌تواند شبکه را از حیطه بی‌درنگ خارج سازد. هدف استفاده از این شبکه آن است که نتایج برداشت شده توسط ریز شبکه عمیق را تصحیح کند. این ریز شبکه از یک لایه پیچشی معمولی به همراه سه لایه پیچشی عمیق تفکیک‌پذیر با اندازه کرنل ۳ × ۳ بهره می‌برد. انتقال کرنل در تمام این لایه‌ها برابر با ۲ در نظر گفته شده به غیر از لایه نهایی که مقدار آن به ۱ تغییر یافته است. در نهایت

[1] Stochastic Gradient Descent

ویژگی‌های دریافتی از هر دو ریز شبکه با یکدیگر جمع ساده گشته و اطلاعات جمع شده وارد یک لایه پیچشی معمولی جهت طبقه‌بندی و اعمال طبقه‌بند softmax می‌شوند.

۳-۳- جزئیات آموزش و کلاس‌های انتخابی

در این بخش به توضیح پارامترهای متفاوت جهت آموزش این مدل‌ها پرداخته می‌شود. به منظور اجتناب از آموزش طولانی مدل‌های یادگیری عمیق، در این بخش تمام تصاویر پایگاه داده‌های آموزش و آزمایش به اندازه ۳۶۰ × ۴۸۰ تبدیل شده است. از کتابخانه Pytorch جهت اعمال فرآیند یادگیری ماشین استفاده گردیده است. جهت آموزش تمامی مدل‌ها از نرخ یادگیری اولیه ۰/۰۰۰۵ استفاده شده است که این نرخ یادگیری در طول فرایند آموزش تغییر می‌کند. اندازه دسته ورودی ۶۴ در نظر گرفته شده است. مطابق با مرجع [۴۹] اندازه دسته ورودی بالا می‌تواند در حین استفاده از بهینه‌سازهایی مانند SGD سبب افت دقت گردد. از طرفی استفاده از اندازه دسته ورودی کوچک نیز باعث بالا رفتن زمان مورد نیاز جهت آموزش شبکه می‌گردد. جهت آموزش تمامی مدل‌ها از بهینه‌ساز adam [۵۰] بهره گرفته شده است. همچنین تابع اتلاف مورد استفاده قرار گرفته نیز اختلاف انتروپی بوده و طبقه‌بند مورد استفاده قرار گرفته نیز softmax است. نحوه سنجش اختلاف میان پیش‌بینی مدل و حقیقت محض متناظر توسط تابع اتلاف اختلاف انتروپی در رابطه ۱ نشان داده شده است.

$$l = -(g_i \log(p_i) + (1 - g_i) \log(1 - p_i)) \qquad (۱)$$

در این رابطه علائم g_i و p_i به ترتیب حقیقت محض و پیش‌بینی مدل از پیکسل i هستند. جهت پیشگیری از بیش- برازش[1] و نیز افزایش قابلیت تعمیم مدل‌ها، بر برخی تصاویر به‌طور تصادفی آینه کردن افقی و نیز ضرایب تبدیل تصویر تصادفی بین ۰/۵ الی ۲ اعمال گشته است. همچنین جهت اطمینان از عدم کم- برازش[2] تمامی مدل‌ها برای ۳۰۰ اپوک آموزش داده شده‌اند.

کلاس‌های انتخابی در این مقاله شامل جاده (آسفالت)، پیاده‌رو، چمن، خودرو، درخت، ساختمان، آسمان و عابرین پیاده است که در مجموع هشت کلاس متفاوت را پوشش می‌دهند.

۴- ارزیابی نتایج

در این بخش عملکرد شبکه‌ها از نظر دقت و سرعت پردازش مورد ارزیابی قرار گرفته و نتایج به‌دست آمده تفصیر خواهند شد. در ابتدا نتایج آموزش شبکه‌های معلم با استفاده از BDD100K بررسی شده و سپس قابلیت تعمیم و مقاوم بودن آن‌ها توسط شبیه‌ساز CARLA سنجیده خواهد شد. در پایان شبکه دانش‌آموز توسط شبکه‌های معلم آموزش داده شده و نتایج آن با زمانی که این آموزش توسط انسان صورت می‌پذیرد مقایسه خواهند شد.

۴-۱- آموزش شبکه‌های معلم و ارزیابی آن‌ها

جدول (۱) ارزیابی مقایسه‌ای از عملکرد شبکه‌های یادگیری عمیق انتخاب شده را برای تمامی کلاس‌ها با معیار mIoU نشان می‌دهد. معیار ارزیابی این مقاله در حیطه قطعه‌بندی محیط mIoU انتخاب گشته است چرا که استفاده از روشی مانند دقت پیکسل به علت عدم توازن در پیکسل‌های کلاس‌ها ممکن است نتایجی غیر قابل تطابق با عملکرد مدل در اختیار بگذارد. رابطه ۲ روش ارزیابی مدل را با استفاده از معیار mIoU نشان می‌دهد.

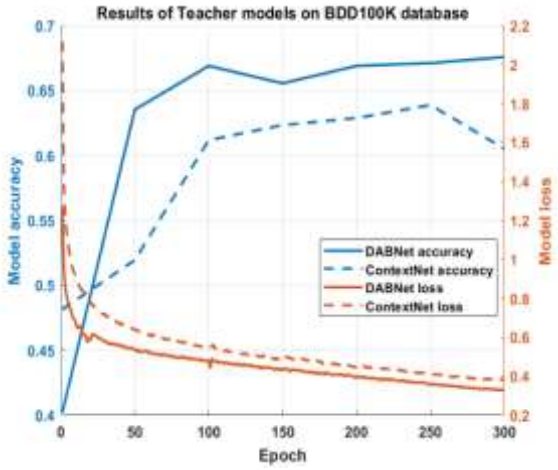
$$mIoU = \frac{TP}{TP + FP + FN} \qquad (۲)$$

در این رابطه، علائم TP، FP و FN به ترتیب بیانگر تعداد پیکسل‌های مثبت صحیح، مثبت کاذب و منفی کاذب هستند. جدول (۱) علاوه بر مقایسه دقت مدل‌های آموزش داده شده، شبکه‌ها را از نظر سرعت پردازشی نیز با یکدیگر مقایسه می‌نماید. سرعت پردازشی گزارش شده در این جدول با استفاده از Nvidia Geforce GTX 1050 مورد آزمایش قرار گرفته است. بهترین دقت برای هر مدل با رنگ آبی و بهترین دقت در مقایسه تمامی مدل‌ها با رنگ قرمز نمایان گشته است. همچنین در شکل (۲) نمودار افزایش دقت و کاهش تابع اتلاف برای هر دو معماری DABNet و ContextNet نمایش داده شده است. نمونه تصاویری از مقایسه بصری مدل‌های آموزش داده شده در شکل (۳) قابل مشاهده است. همانطور که در نمودار شکل (۲) و همچنین جدول (۱) مشاهده می‌شود، DABNet در مقایسه با ContextNet از نقطه نظر دقت، عملکرد بسیار مناسب‌تری را از خود نشان داده است. اما این عملکرد بهتر منجر به اندکی کاهش سرعت پردازشی در مقایسه با ContextNet شده است. همانطور که در قسمت‌های قبل نیز ذکر شد، سرعت پردازشی بالا اولویت شبکه معلم نبوده و در مقایسه انجام شده، DABNet به‌دلیل عملکرد بسیار مناسبتر از نظر دقت، برای جایگاه معلم در اولویت قرار می‌گیرد.

در شکل (۲) پس عبور از ۲۵۰ ایپوک، دقت ContextNet بر پایگاه داده آزمایش با افت عملکرد روبرو شده است. علت این امر آن است که ContextNet از یک ریز شبکه با عمق بسیار کم

[1] Overfitting
[2] Underfitting

استفاده می‌کند که به جای یادگیری ویژگی‌های استخراج شده اقدام به حفظ آن‌ها می‌کند. این امر موجب شده است که شبکه مذکور پس از ۲۵۰ ایپوک با بیش- برازش مواجه شود و با وجود افزایش دقت در پایگاه آموزش، در داده‌های آزمایش با افت عملکرد چشم‌گیری روبرو گردد. این در حالی است که ماژول‌های DAB در DABNet به علت بهره‌گیری از ساختار عمیق الهام گرفته از ResNet ویژگی‌های تصویر را یاد گرفته و توانسته است دقت مدل را برای داده‌های آزمایش نیز بالا نگه دارد.

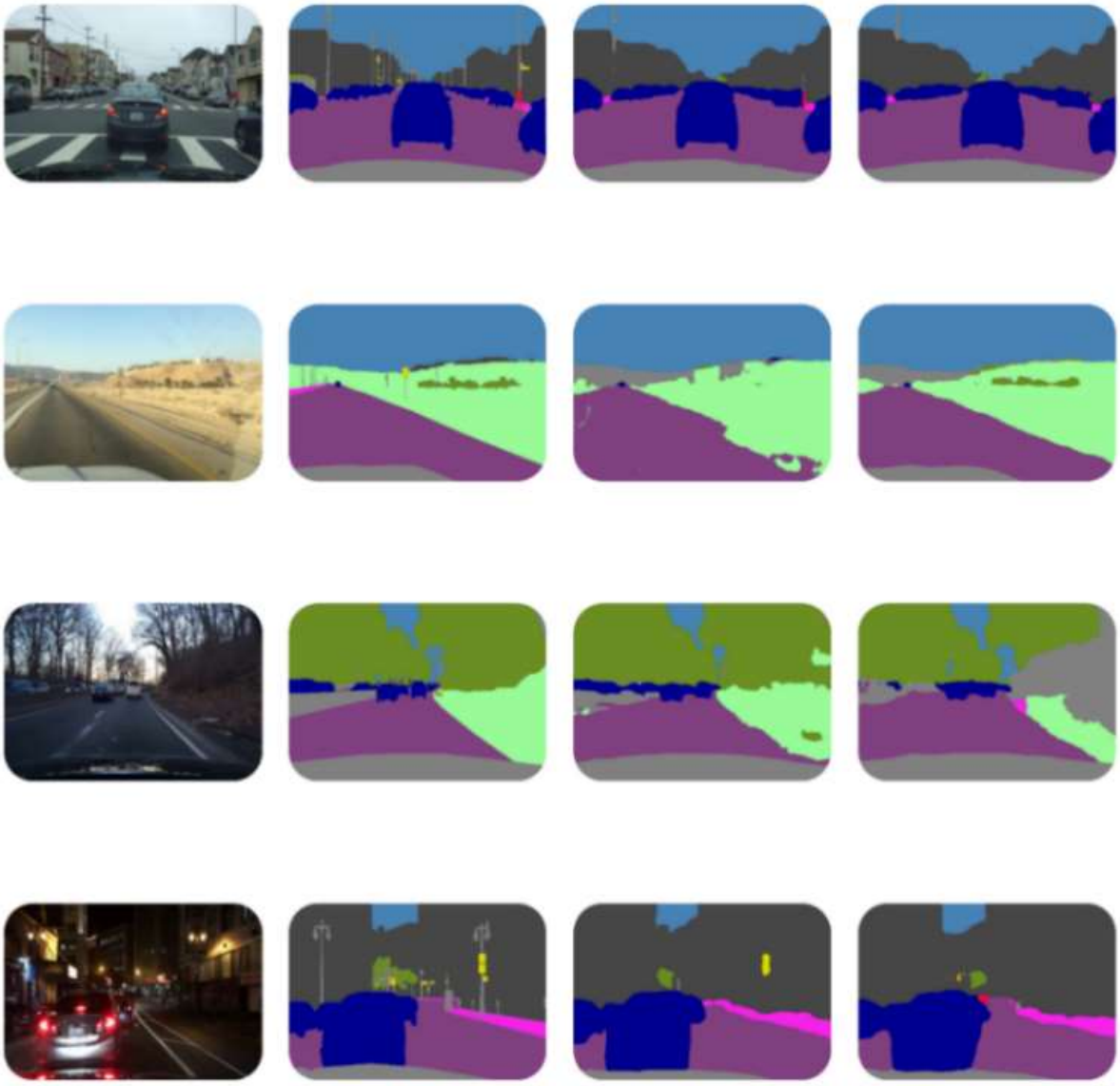


شکل (۲): نمودار افزایش دقت و کاهش تابع اتلاف شبکه‌های معلم

در ارزیابی شبکه‌های معلم عملکرد بهتر مدل DABNet در حالی است که از نقطه نظر تعداد پارامترها این شبکه تعداد پارامتر کمتری (۰/۷۶ میلیون) را در اختیار دارد. در عموم شرایط این گونه برداشت می‌شود که هر چه تعداد پارامترهای شبکه بیشتر باشد و شبکه به سمت عمیق‌تر بودن برود، توانایی آن در استخراج ویژگی‌های پیچیده افزایش یافته و دقت آن بهبود خواهد یافت. اما باید توجه داشت که مدل‌های بزرگ گاهی تعداد بسیار زیادی پارامتر زائد در معماری خود دارند. به‌عنوان مثال ICNet مرجع [۴۸] با وجود اینکه بیش از ۱۰ برابر DABNet پارامتر در اختیار دارد، اما با این حال از نقطه نظر دقت از DABNet ضعیف‌تر عمل می‌نماید [۴۰]. در نقطه مقابل اما علت سرعت پردازشی بالاتر ContextNet را می‌توان در استخوان‌بندی آن جویا شد. به‌طور کلی استخوان‌بندی این شبکه مشابه MobileNet V2 [۴۷] است که سرعت پردازشی بالاتری را در مقایسه با شبکه‌ای با استخوان‌بندی مشابه با ResNet از خود نشان می‌دهد. البته استفاده از ساختار مشابه با ResNet، DABNet را به سمت عمیق‌تر رفتن برده و انتظار اختلاف بسیار زیادی پارامتر زائد در معماری خود دارند. به‌عنوان مثال ICNet مرجع [۴۸] با وجود اینکه بیش از ۱۰ برابر DABNet پارامتر در اختیار دارد، اما با این حال از نقطه نظر دقت از DABNet ضعیف‌تر عمل می‌نماید [۴۰]. در نقطه مقابل اما علت سرعت پردازشی بالاتر ContextNet را می‌توان در استخوان‌بندی آن جویا شد. به‌طور کلی استخوان‌بندی این شبکه مشابه MobileNet V2 [۴۷] است که سرعت پردازشی بالاتری را در مقایسه با شبکه‌ای با استخوان‌بندی مشابه با ResNet از خود نشان می‌دهد. البته استفاده از ساختار مشابه با ResNet، DABNet را به سمت عمیق‌تر رفتن برده و انتظار اختلاف چشم‌گیرتر در سرعت پردازشی می‌رود. با این‌حال همانطور که در نتایج جدول (۱) نیز مشاهده می‌شود این شبکه چندان در مقایسه با ContextNet با کاهش سرعت پردازشی مواجه نیست (۴/۲ فریم بر ثانیه ضعیف‌تر از ContextNet). علت این امر حذف بخش رمزگشا در معماری DABNet است. حذف رمزگشا اگرچه منجر به کاهش دقت می‌گردد اما سرعت پردازشی را به مراتب افزایش بخشیده و به همین دلیل اختلاف زمان پردازش میان ContextNet و DABNet عدد بزرگی نیست.

جدول (۱): نتایج ارزیابی شبکه‌های معلم در پایگاه داده BDD100K

مدل	کلاس								FPS	پارامتر
	جاده	پیاده‌رو	چمن	خودرو	عابر	آسمان	درخت	ساختمان		
DABNet	۸۷/۸	۵۱/۲	۳۸/۷	۸۳/۱	۵۴/۹	۹۱/۷	۸۰/۱	۷۵/۷	۷۹/۱	۰/۷۶
ContextNet	۸۰/۰	۴۶/۵	۳۲/۵	۷۷/۲	۲۰/۵	۸۰/۹	۷۵/۲	۷۱/۴	۸۴/۲	۰/۸۸

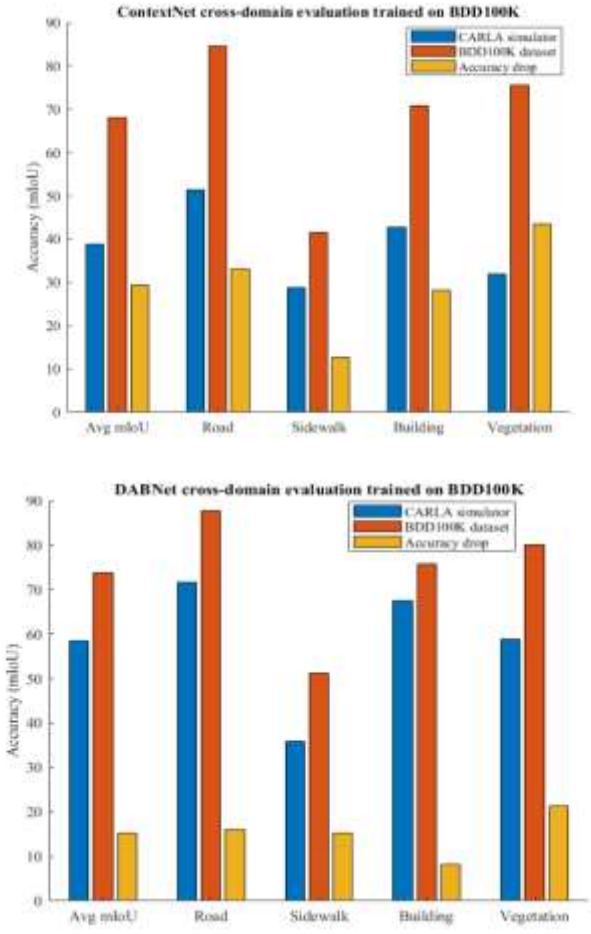


شکل (۳): تصاویری از قطعه‌بندی انجام شده توسط شبکه‌های معلم در پایگاه داده BDD100K. از سمت چپ به راست ستون اول: تصاویر ورودی، ستون دوم: حقیقت محض تصویر، ستون سوم: قطعه‌بندی انجام شده توسط DABNet و ستون چهارم: قطعه‌بندی انجام شده توسط ContextNet

۴-۲- ارزیابی شبکه‌های معلم با دامنه متفاوت

همانطور که در بخش‌های پیشین ذکر شد، شبکه معلم باید از قابلیت تعمیم خوبی برخوردار باشد تا بتوان جهت آموزش یک شبکه دانش‌آموز در دامنه‌ای متفاوت به آن استناد نمود. در این بخش هدف آن است که مقاوم بودن مدل‌های آموزش داده شده با تصاویری بسیار چالشی که اختلاف دامنه شدیدی با داده‌های آموزش داده شده دارند مورد بررسی قرار گیرد. بدین منظور از شبیه‌ساز CARLA [۵۱] جهت شبیه‌سازی تصاویر بسیار چالشی و مصنوعی در شهرها، شرایط آب و هوایی و همچنین شدت نور متفاوت استفاده شده است. این شبیه‌ساز همزمان با تولید تصاویر شبیه‌سازی، برچسب‌گذاری متناظر با هر تصویر را نیز با رعایت استاندارد Cityscapes در اختیار قرار می‌دهد.

شکل (۴): نمونه تصاویری از شبیه‌سازی انجام شده در CARLA



شکل (۵): نمودار میله‌ای ارزیابی مقاوم بودن مدل‌های معلم و افت عملکرد آن‌ها توسط اختلاف دامنه شدید موجود در شبیه‌ساز CARLA

جدول (۲): نتایج ارزیابی با دامنه متفاوت شبکه‌های قطعه‌بندی در شبیه‌ساز CARLA.

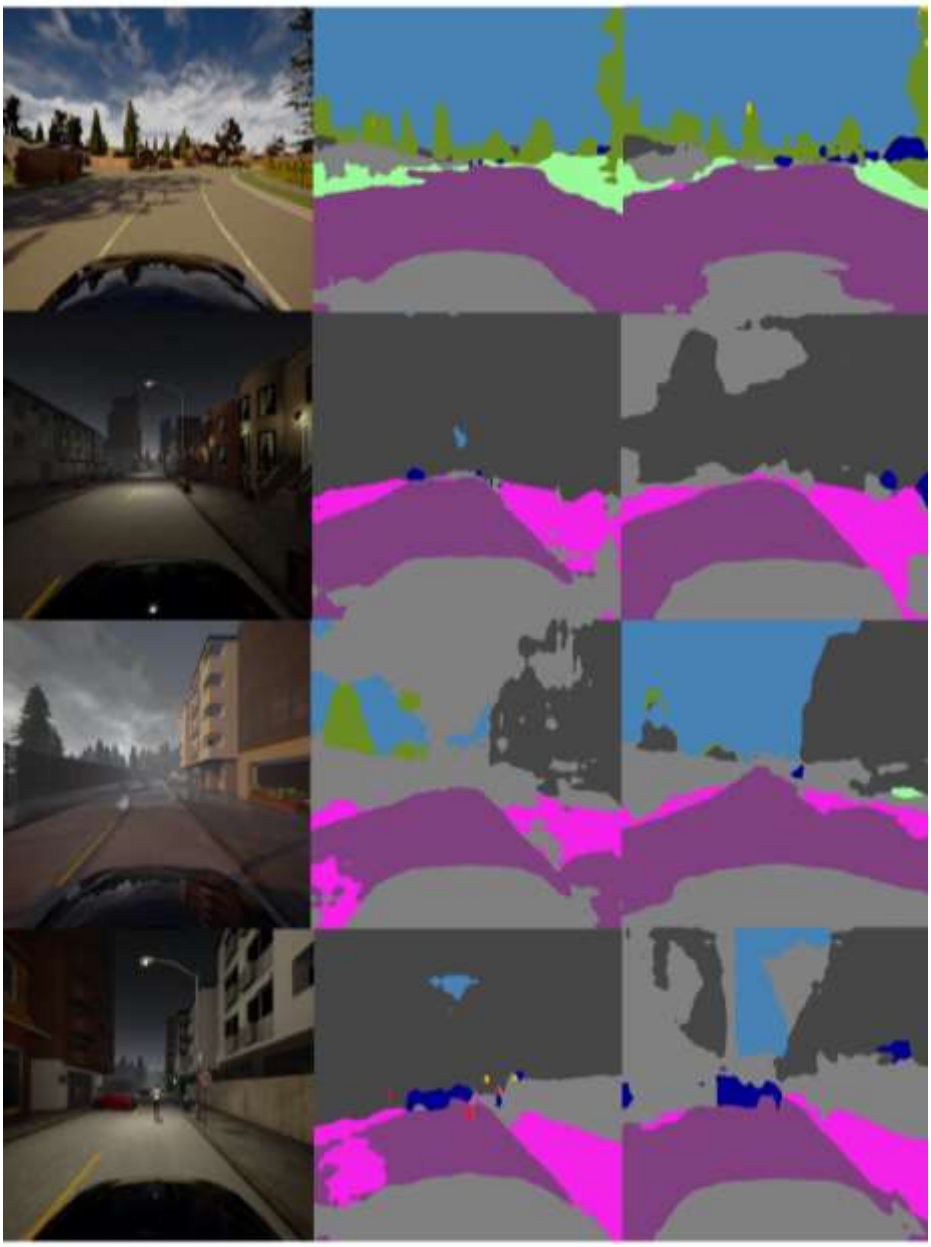
مدل	جاده	پیاده‌رو	ساختمان	درخت	دقت کل
DABNet	۷۱/۷	۳۵/۹	۶۷/۵	۵۸/۸	۵۸/۵
ContextNet	۵۱/۴	۲۸/۸	۴۲/۷	۳۲	۳۸/۷
FastSCNN	۶۸/۹	۳۱/۷	۵۴/۵	۴۸/۶	۵۰/۹

وجه تمایز اصلی استفاده از شبیه‌ساز CARLA در مقایسه با تصاویر واقعی BDD100K را می‌توان در تفاوت شهر، مسیر حرکت و شرایط استاتیکی محیط شبیه‌سازی شده توصیف کرد. به همین علت در ارزیابی مقاوم بودن مدل‌ها تنها کلاس‌های جاده، پیاده‌رو، ساختمان و درختان مورد بررسی قرار گرفته‌اند. ارزیابی کلاس چمن در این قسمت مقدور نیست چرا که شبیه‌ساز CARLA از کلاس چمن پشتیبانی نمی‌کند. شکل (۴) نمونه تصاویری از شبیه‌سازی انجام شده در CARLA را نمایش می‌دهد. نتیجه ارزیابی مقاوم بودن مدل‌های آموزش داده شده در جدول (۲) گزارش شده است. همچنین شکل ۵ مقایسه دقت مدل‌های معلم را برای BDD100K و CARLA به شکل نمودار میله‌ای به نمایش گذاشته است. جهت انجام این ارزیابی ۱۹۴۲ تصویر در نرم‌افزار CARLA شبیه‌سازی شده است که این عدد در حدود دو برابر پایگاه داده آزمایش BDD100K است. جهت اطمینان از اختلاف دامنه شدید، محوریت شبیه‌سازی بر شرایط آب و هوایی بارانی و شب بوده است. شکل (۶) نمونه تصاویری از مقاوم بودن مدل‌های معلم را در تصاویر استخراج شده از CARLA نمایش می‌دهد. مطابق انتظار، DABNet در ارزیابی با دامنه متفاوت نیز عملکرد بهتری را از خود نشان داده است که نشان می‌دهد این شبکه در مقایسه با ContextNet علاوه بر به‌دست آورد دقت بالاتر، از قابلیت تعمیم و مقاومت بالاتری نیز برخوردار است. از نقطه نظر افت دقت نیز مطابق با شکل ۵، DABNet با میانگین افت ۱۵/۲۲٪ مقاومت بهتری را در محیط جدید و متفاوت CARLA از خود نشان داده است. این عدد در ContextNet به عدد ۲۹/۴٪ افزایش میابد که به‌طور تقریبی منجر به دو برابر افت عملکرد بیشتر نسبت به DABNet شده است. علت این امر آن است که مبتنی بودن معماری به ریز شبکه کم عمق باعث می‌شود جزئیات برداشت شده هنگام استخراج ویژگی همواره در راستای دامنه محیط آموزش باشد و در شرایطی که اطلاعات محیطی تصویر معرف نوع جدیدی از دامنه محیط باشد، ریز شبکه عمیق ContextNet در درک محیط موفق عمل نکرده و در نتیجه قطعه‌بندی حاصل با افت شدیدی مواجه خواهد بود. این در حالی است که ماژول DAB در DABNet توانسته است از یادگیری بر مبنای بافت تصویر اجتناب کرده و متمرکز بر شکل هر کلاس شود. در نتیجه با عوض شدن کامل دامنه تصاویر (که منجر به تغییر بافت تصویر می‌شود) اگرچه این شبکه نیز مطابق انتظار با افت دقت مواجه خواهد شد اما این افت به مراتب از شبکه‌ای نظیر ContextNet کمتر بوده است.

از نقطه نظر مقایسه کلاس‌ها به صورت تنها نیز کلاس جاده بهترین دقت را در تمام شبکه‌ها به خود اختصاص داده است. این نتیجه کامل بدیهی است، چون در پایگاه داده آموزش BDD100K جاده همواره نزدیک‌ترین و گسترده‌ترین کلاس پیش روی خودرو است. بنابراین استخراج ویژگی از تصاویر برای کلاس جاده همواره با بهترین نتیجه همراه بوده است. از طرفی، در شبیه‌سازی CARLA نیز جاده اولین کلاس پیش روی خودرو است که منجر به دریافت بالاترین دقت در مقایسه با سایر کلاس‌ها شده است. کلاس پیاده‌رو از نظر ارزیابی با چالش بزرگی مواجه است. علت این امر آن است که هنگام آموزش شبکه‌ها علاوه بر کلاس پیاده‌رو به آموزش کلاس چمن نیز پرداخته شده است. در حالی که شبیه‌سازی CARLA از کلاس چمن پشتیبانی نکرده و تمامی پیکسل‌هایی که حاوی چمن است را در کلاس پیاده‌رو طبقه‌بندی می‌کند. این در حالی است که شبکه‌های آموزش داده شده هنگام مواجه با این تفاصیر به‌دلیل پشتیبانی از کلاس چمن، این پیکسل‌های مذکور را به‌عنوان چمن طبقه‌بندی می‌کند و اگرچه طبقه‌بندی به صورت درست انجام شده است، اما به دلیل تناقض با حقیقت محض تصویر برچسب‌گذاری شده، قطعه‌بندی این پیکسل‌ها خطا برداشت شده و از دقت شبکه‌ها کسر می‌کند (ردیف نخست شکل ۶).

۴-۳- آموزش و ارزیابی شبکه دانش‌آموز

در این بخش از مقاله هدف آن است که با استفاده از شبکه معلم، شبکه یادگیری عمیق دانش‌آموز را بدون استفاده از برچسب‌گذاری دستی، بر پایگاه داده Cityscapes آموزش داده و نتایج آن با هنگامی که آموزش این شبکه مستقیم با استفاده از برچسب‌گذاری دستی انجام می‌شود، مقایسه شود. همانطور که در پیش نیز ذکر شد، هر چقدر شبکه معلم دقت بالاتری را داشته باشد، نتیجه آموزش دانش‌آموز نیز بهتر خواهد بود. به همین دلیل در ابتدا آموزش می‌توان انتظار داشت که شبکه‌ای که توسط DABNet تعلیم دیده است دقت بهتری را از خود نشان بدهد. جهت سادگی در ارجاع به اشکال و جداول، این بخش از مقاله به سه سناریو تقسیم‌بندی شده است.



شکل (۶): نمونه تصاویری از قطعه‌بندی انجام شده در شبیه‌ساز CARLA. از سمت چپ به راست ستون اول: تصاویر شبیه‌سازی شده در CARLA، ستون دوم: قطعه‌بندی خروجی مدل DABNet، ستون سوم: قطعه‌بندی شبکه ContextNet

• **سناریو ۱ (sc1):** در این سناریو به‌طور کلی شبکه‌ای تحت عنوان معلم وجود ندارد. FastSCNN بر تصاویر برچسب‌گذاری دقیق پایگاه داده Cityscapes آموزش داده شده و مورد آزمایش و بررسی قرار می‌گیرد. از این سناریو جهت صحه‌گذاری بر روش معلم- دانش‌آموز استفاده شده است چرا که روش مذکور در نهایت باید پاسخی نزدیک به دقت به‌دست آمده در این سناریو را به‌دست آورد.

• **سناریو ۲ (sc2):** در این سناریو فرض بر این است که هر دو بخش تصاویر آموزش و آزمایش پایگاه داده فاقد برچسب‌گذاری هستند. بنابراین معلم برچسب‌گذاری خودکار را بر تصاویر هر دو بخش پایگاه داده اعمال کرده و شبکه دانش‌آموز با استفاده از آن‌ها آموزش داده شده و مورد ارزیابی قرار می‌گیرد.

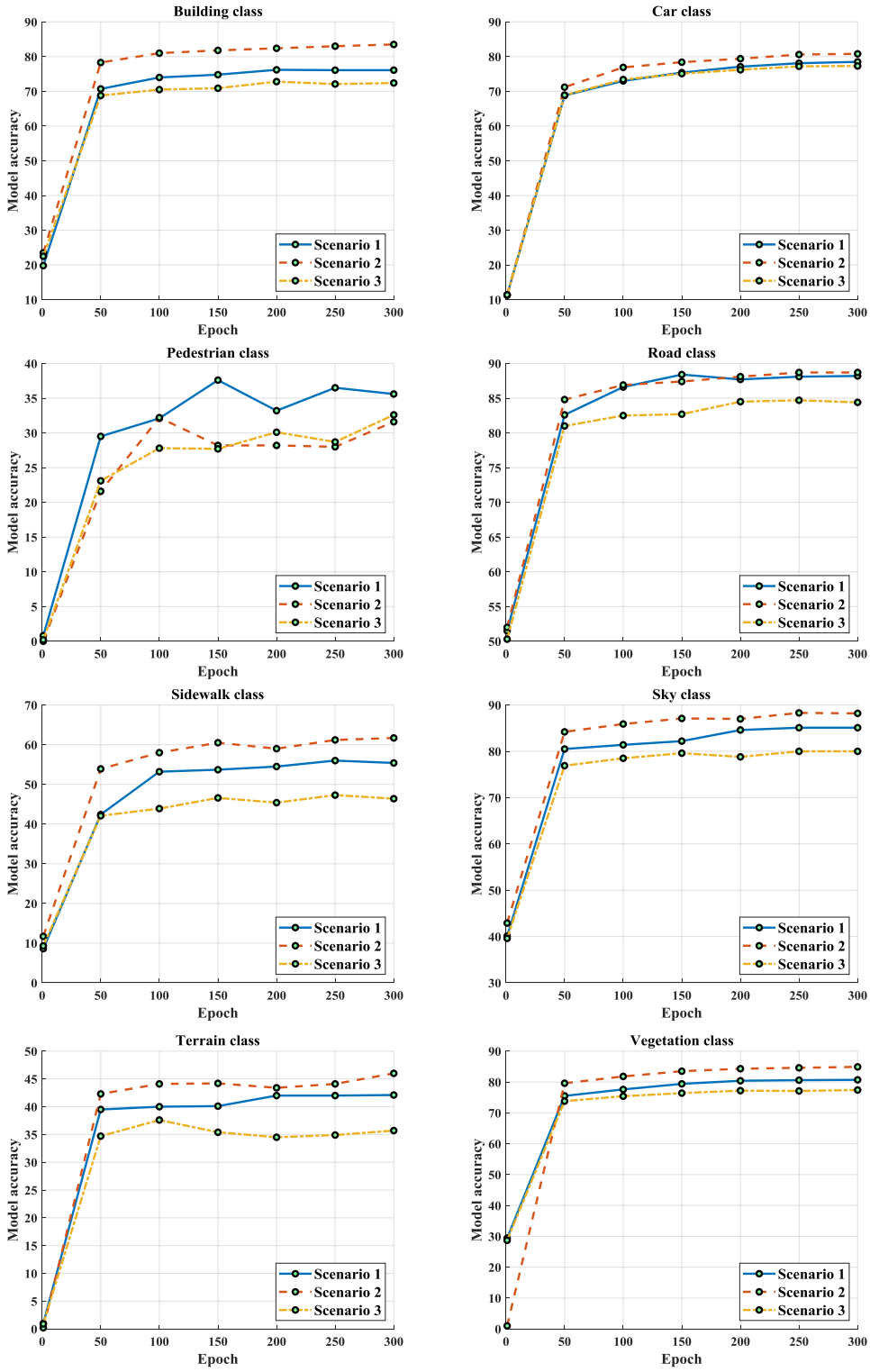
• **سناریو ۳ (sc3):** در تعریف انجام شده در سناریو دوم یک مشکل اساسی وجود دارد و آن نیز این است که تصاویر آزمایش نباید دارای برچسب‌گذاری خودکار شبکه معلم باشند. در واقع، این تصاویر باید به‌طور قاطع توسط برچسب‌گذاری دستی علامت‌گذاری شده باشند چرا که در غیر این‌صورت ارزیابی انجام شده و دقت‌های محاسبه شده مطابق با حقیقت محض تصویر نخواهد بود. در سناریو سوم فقط تصاویر آموزشی توسط معلم برچسب‌گذاری خودکار خواهند گرفت و شبکه دانش‌آموز با

تصاویر آزمایشی که دارای برچسب‌گذاری دستی هستند مورد ارزیابی قرار خواهد گرفت.

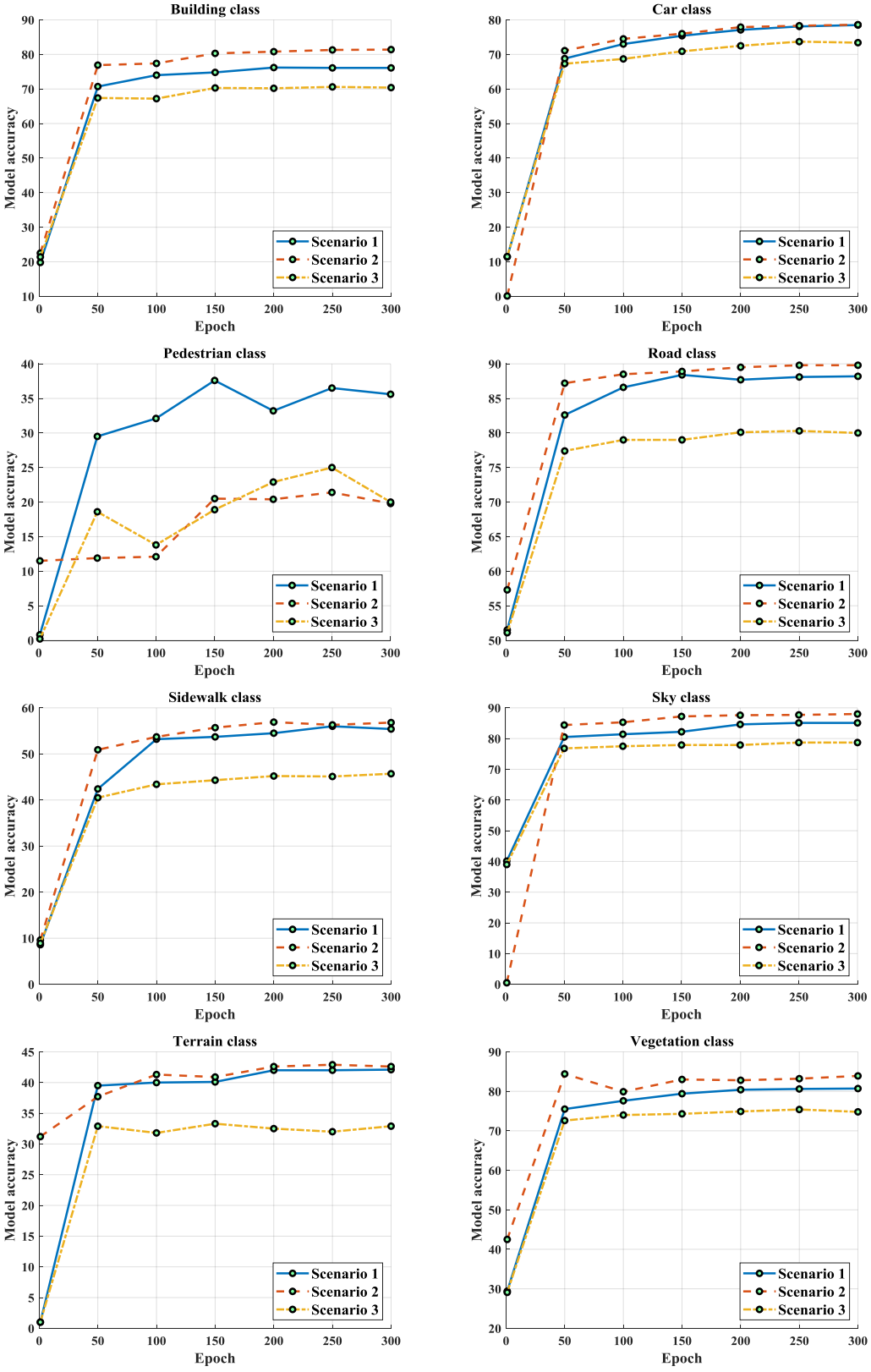
با توجه به این‌که BDD100K در آمریکا و Cityscapes در آلمان جمع‌آوری شده است، بنابراین تغییرات دامنه قابل توجهی در حین انجام برچسب‌گذاری خودکار پیش روی شبکه معلم قرار خواهد گرفت که این امر قابلیت تعمیم این شبکه‌ها را نیز به چالش خواهد کشید. شکل‌های (۷ و ۸) به‌ترتیب نتایج عملکرد شبکه دانش‌آموز را هنگام آموزش توسط DABNet و ContextNet نمایش داده است. همچنین جدول (۳) دقت کل شبکه را در هر سه سناریو تعریف شده با یکدیگر مقایسه می‌نماید. در نهایت شکل (۹) نیز نمونه تصاویری از عملکرد مدل دانش‌آموز نهایی در پایگاه داده آزمایش Cityscapes ارائه نموده است.

همانطور که در جدول (۳) مشاهده می‌شود، مطابق انتظار DABNet معلم بهتری در مقایسه با ContextNet بوده است. FastSCNN تعلیم دیده توسط DABNet بهبود عملکردی معادل با ۳/۶٪ در مقایسه با شبکه مشابه تعلیم دیده توسط ContextNet داشته است. جهت منصفانه بودن آزمایش‌ها، هیچکدام از شبکه‌های آموزش داده شده دارای پیش- آموزش نبوده‌اند. از نقطه نظر درصد اختلاف نیز بهبود ۵/۷٪ شبکه تعلیم داده شده توسط DABNet در مقایسه با ContextNet نشان می‌دهد که برچسب‌گذاری اتوماتیک توسط DABNet به مراتب به برچسب‌گذاری دستی نزدیک‌تر و قابل استنادتر است. اعداد فوق همچنین ثابت می‌کنند که دقت دانش‌آموز همیشه به تعداد پارامترهای شبکه معلم بستگی ندارد بلکه نکته مهم موثر بودن این پارامترها در استخراج ویژگی‌های پیچیده محیط است. اگرچه با فرض اینکه اکثریت پارامترهای شبکه معلم موثر هستند، می‌توان ادعا نمود که هرچقدر مدل معلم عمیق‌تر باشد، دقت دانش‌آموز به دقت برچسب‌گذاری دستی نزدیک‌تر خواهد بود. با وجود استفاده از تصاویر با اندازه‌های کوچک و نیز انتخاب شبکه‌ای بی‌درنگ به‌عنوان معلم، دانش‌آموز (FastSCNN) توانسته است ۹۳٪ دقت به‌دست آمده در برچسب‌گذاری دستی را با تکیه کامل بر روش برچسب‌گذاری خودکار با تعلیم دیدن از DABNet به‌دست آورد. نتایج سناریو ۲ در جدول (۳) نشان می‌دهد که دقت دانش‌آموز هنگامی که شبکه معلم بر پایگاه داده آزمایش نیز برچسب‌گذاری خودکار اعمال کند به مراتب بالاتر از دقت در فرم حقیقت محض تصویر (سناریو ۳) است. این اتفاق به دلیل آن است که شبکه‌های هوش مصنوعی گاهی پیش‌بینی یکسانی برای برخی پیکسل‌ها دارند. گاهی حتی اگر این پیش‌بینی اشتباه هم باشد، شبکه‌های متعددی این اشتباه را تکرار می‌کنند. این پدیده سبب شده است که شبکه معلم برای برخی پیکسل‌های تصاویر پایگاه داده آزمایش پیش‌بینی اشتباهی انجام دهد. هنگامی که شبکه دانش‌آموز نیز اشتباه مشابهی را برای همان پیکسل انجام می‌دهد و طبقه‌بند دانش‌آموز همان برچسب اشتباه یکسان طبقه‌بند معلم را به پیکسل نسبت بدهد، از آن‌جا که تصویر پیش‌بینی دانش‌آموز به جای حقیقت محض با تصویر پیش‌بینی معلم سنجیده می‌شود، پیکسل مورد نظر به اشتباه یک مثبت حقیقی شناخته شده و منجر به افزایش دقت سناریو ۲ می‌گردد. بنابراین توصیه می‌شود هنگام استفاده از روش معلم- دانش‌آموز، تصاویر آزمایش حتما دارای برچسب‌گذاری دستی (حقیقت محض) باشند تا بتوان ارزیابی صحیحی از دقت به‌دست آمده شبکه دانش‌آموز به‌دست آورد. پدیده مذکور خود را در نمودارهای تابع اتلاف نیز نشان داده است. همانطور که در این شکل مشاهده می‌شود، تابع اتلافی شبکه دانش‌آموز هنگامی که با برچسب‌گذاری خودکار معلم آموزش می‌بیند همواره کمتر از شبکه متناظری است که آموزش آن با برچسب‌گذاری دستی صورت می‌پذیرد. این اتفاق نیز همانگونه که توضیح داده شد به این دلیل است که هنگام آموزش دیدن یک هوش مصنوعی توسط یک هوش مصنوعی دیگر، تکرار خطاها بین معلم و دانش‌آموز می‌تواند سبب یافت نشدن خطا شده و در نتیجه تابع اتلاف همواره مقادیر کمتری را به خود اختصاص می‌دهد. در خصوص تفسیر نتایج هر یک از کلاس‌ها، نتایج بار دیگر تأثیر بزرگ و فراوان بودن یک کلاس خاص در پایگاه داده آموزش معلم را پررنگ می‌نمایند. همانطور که در این نمودارهای و جداول مشاهده می‌شود، در کلاس‌های که پیکسل‌های زیادی از پایگاه داده را به خود اختصاص داده‌اند اختلاف دقت سناریو ۱ با ۳ (استفاده یا عدم استفاده از روش معلم- دانش‌آموز) بسیار کاهش یافته که این امر به استفاده از برچسب‌گذاری خودکار در این کلاس‌ها اعتبار می‌بخشد. کلاس‌های بزرگ را در آزمایش‌های انجام شده می‌توان جاده، ساختمان، درختان، آسمان و خودرو تعریف نمود. افت دقت برچسب‌گذاری خودکار در این کلاس‌ها [۳/۸، ۳/۷، ۳/۳، ۵/۱، ۱/۲] و [۸/۲، ۵/۷، ۵/۹، ۶/۴، ۵/۱] به ترتیب برای شبکه‌های معلم DABNet و ContextNet به‌دست آمده‌اند. شرایط مذکور برای کلاس فراوان اما کوچکی نظیر اشخاص متفاوت است. در کلاس اشخاص شبکه تعلیم داده شده توسط ContextNet با افت بسیار زیاد ۱۵/۶٪ مواجه شده است در حالی که شبکه متناظر با تعلیم DABNet توانسته است این افت را در ۳٪ مهار کند. علت این امر آن است که کلاس اشخاص به دلیل نوع شکل‌گیری دست‌ها، پاها، نحوه پوشش و لباس‌های مختلف و فاصله متفاوت اندام از یکدیگر به‌طور رایج دارای تنوع زیاد در امر قطعه‌بندی محیط است. توانایی ماژول DAB در مدل معلم DABNet جهت شناسایی بهتر پیکسل‌های مرتبط با اشخاص باعث شده است تا

اختلاف میان برچسب‌گذاری دستی و خودکار در این کلاس شدید نباشد. در حالی استفاده ContextNet از ریز شبکه کم عمق باعث ضعف این شبکه به‌عنوان معلم جهت آموزشی صحیح به FastSCNN شده است.



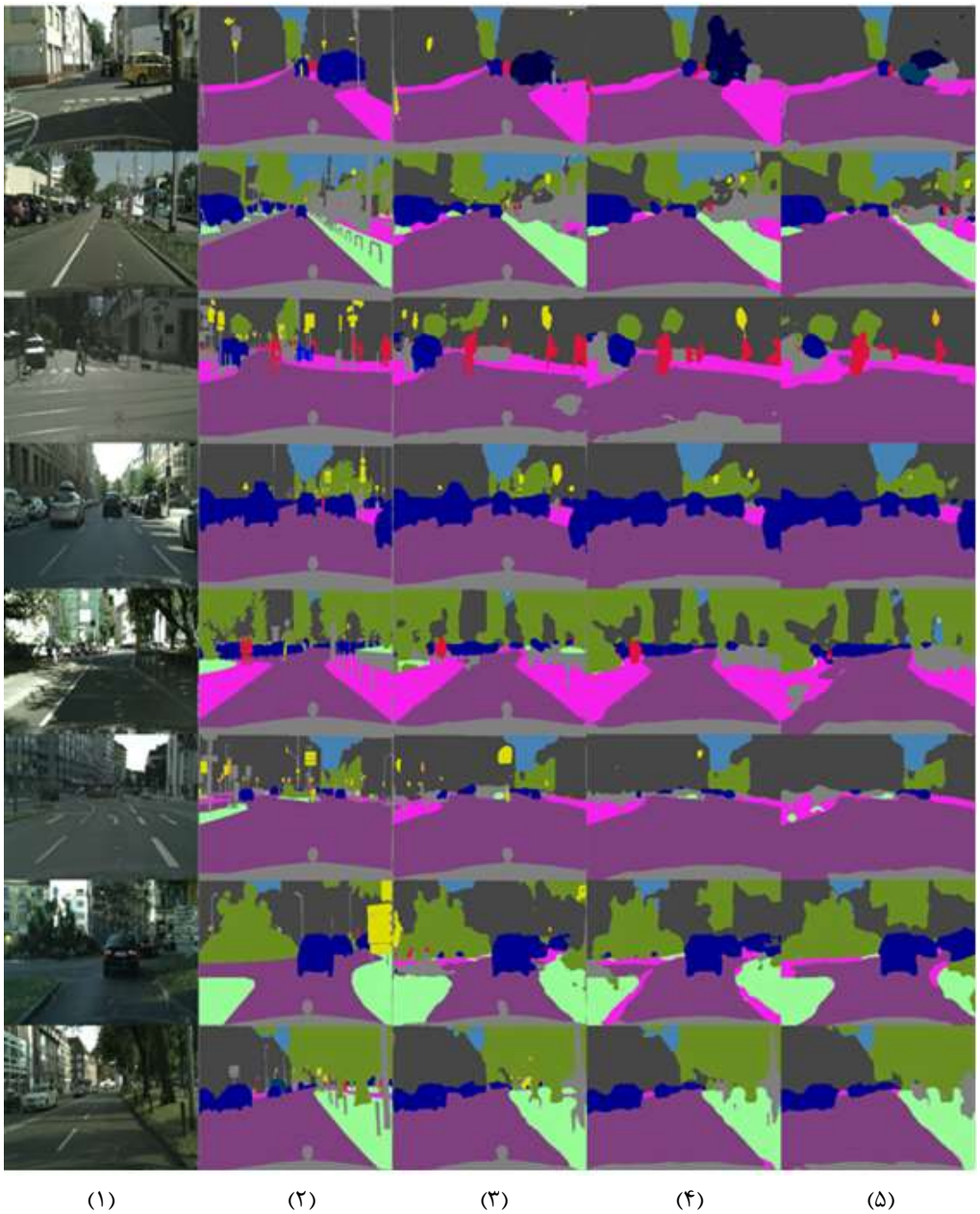
شکل (۷). نمودارهای ارزیابی دقت‌های حاصل شده در آموزش شبکه FastSCNN (دانش‌آموز) با استفاده از DABNet به‌عنوان معلم.



شکل (۸): نمودارهای ارزیابی دقت‌های حاصل شده در آموزش شبکه FastSCNN (دانش‌آموز) با استفاده از ContextNet به‌عنوان معلم.

جدول (۴) مقایسه میان نتایج به‌دست آمده در این پژوهش و سایر مقالات مرتبط با این حوزه را نمایش می‌دهد. همانطور که قابل مشاهده است روش معلم-دانش‌آموز مورد استفاده توانسته با وجود استفاده از اندازه تصاویر کوچک‌تر در پایگاه داده خود (که

منجر به از دست رفتن حجم داده قابل توجهی شده است) به دقتی قابل رقابت با سایر پژوهش‌ها برسد. همچنین لازم به ذکر است که دقت مقایسه شده در این جدول هنگام استفاده از تعداد داده برابر نیز قابل رقابت با سایر مقالات است. این در حالی است که روش پیشنهادی به راحتی قادر است انبوهی از داده را برای فرایند آموزش شبکه مورد استفاده قرار دهد و محدود به داده‌های برچسب‌گذاری شده انسانی نیست. شبکه‌های معلم مورد استفاده، هر دو بلادرنگ هستند که می‌توان با بهره بردن از شبکه‌های عمیق‌تر دقت این رویکرد را بهبود بخشید. با وجود شرایط ذکر شده، شبکه آموزش داده شده توسط DABNet توانسته است به ترتیب ۱/۶٪ و ۳/۸٪ بهتر از شبکه‌های ENet و SegNet عمل نموده و دقتی در نزدیکی ESPNet را با در اختیار داشتن یک سوم رزولوشن کوچکتر تصاویر به ثبت رساند. نتایج به دست آمده نشان می‌دهد روش کاری مورد استفاده در این مقاله به خوبی قادر است خود را با دامنه‌های جدید تصاویر تطبیق داده و از افت عملکرد شبکه دانش‌آموز جلوگیری نماید.



(۱) (۲) (۳) (۴) (۵)

شکل (۹): نمونه تصاویر قطعه‌بندی انجام شده در پایگاه داده آزمایش Cityscapes توسط شبکه FastSCNN. (۱): تصویر ورودی، (۲): حقیقت محض، (۳): آموزش توسط انسان، (۴): آموزش توسط DABNet و (۵): آموزش توسط ContextNet.

جدول (۳): ارزیابی مقایسه‌ای روش معلم- دانش‌آموز در هر سه سناریو.

معلم	Epoch	دقت			تمایز دقت سناریو ۱ و ۳	
		sc1	sc2	sc3	اختلاف	درصد اختلاف
DABNet	۱	۱۸/۲	۱۶/۰	۱۸/۲	۰	۰
	۵۰	۵۷/۷	۶۱/۷	۵۵/۱	۲/۶	۴/۵
	۱۰۰	۶۱/۴	۶۵/۹	۵۷/۷	۳/۷	۶
	۱۵۰	۶۲/۹	۶۶/۳	۵۸/۳	۴/۶	۷/۳
	۲۰۰	۶۳/۵	۶۶/۶	۵۹/۰	۴/۵	۷/۱
	۲۵۰	۶۴/۵	۶۷/۹	۵۹/۵	۵	۷/۸
	۳۰۰	۶۴/۴	۶۸/۶	۵۹/۹	۴/۵	۶/۹
ContextNet	۱	۱۸/۲	۱۹/۵	۱۸/۱	۰/۱	۰/۵
	۵۰	۵۷/۷	۶۰/۲	۵۳/۲	۴/۵	۷/۸
	۱۰۰	۶۱/۴	۶۱/۴	۵۳/۵	۷/۹	۱۲/۹
	۱۵۰	۶۲/۹	۶۴/۰	۵۵/۵	۷/۴	۱۱/۸
	۲۰۰	۶۳/۵	۶۴/۷	۵۶/۳	۷/۲	۱۱/۳
	۲۵۰	۶۴/۵	۶۵/۰	۵۶/۸	۷/۷	۱۱/۹
	۳۰۰	۶۴/۴	۶۵/۱	۵۶/۳	۸/۱	۱۲/۶

جدول (۴): مقایسه نتایج به دست آمده در این پژوهش با کارهای سایر مقالات.

مدل	رویکرد	اندازه تصویر	دقت	مرجع
SegNet	نظارت شده	۱۰۲۴ × ۲۰۴۸	۵۶/۱	[۵۲]
ENet	نظارت شده	۱۰۲۴ × ۲۰۴۸	۵۸/۳	[۵۳]
FPENet	نظارت شده	۳۸۴ × ۷۶۸	۶۲/۷	[۵۴]
ESPNet	نظارت شده	۵۱۲ × ۱۰۲۴	۶۰/۳	[۵۵]
ESPNetV2	نظارت شده	۵۱۲ × ۱۰۲۴	۶۲/۱	[۵۶]
معلم DABNet	نیمه-نظارت شده	۳۶۰ × ۴۸۰	۵۹/۹	این مقاله
معلم ContextNet	نیمه-نظارت شده	۳۶۰ × ۴۸۰	۵۶/۳	این مقاله

۵- جمع‌بندی

در این مقاله با برقرای ارتباط بین مدل‌های یادگیری عمیق توسط روش معلم- دانش‌آموز، قطعه‌بندی معنایی روی پایگاه‌های داده خودروهای خودران انجام شد. رویکرد این مقاله برخلاف رویکردهای نظارتی نه تنها به سادگی می‌تواند از انبوه داده‌های فاقد برچسب بهره ببرد، بلکه در مقایسه با رویکرد نظارتی به زمان بسیار کوتاه‌تری جهت برچسب‌گذاری پایگاه داده‌ای جدید نیاز دارد. معماری‌های DABNet و ContextNet که با استفاده از پایگاه داده BDD100K آموزش داده شده‌اند، پایگاه داده Cityscapes را به‌طور کامل و بدون دخالت انسان در فرآیند یادگیری به مدل FastSCNN آموزش داده‌اند. همچنین با توجه به این که شبکه‌های معلم به‌طور رایج در وضعیت ارزیابی مقاوم بودن قرار دارند، بنابراین قابلیت تعمیم این شبکه‌ها نیاز به بررسی دارد. این بررسی با شبیه‌سازی تصاویر مصنوعی با اختلاف دامنه شدید نسبت به پایگاه داده آموزش BDD100K در شبیه‌ساز CARLA صورت گرفته است. نتایج نشان می‌دهد با وجود جلوگیری از اتلاف زمان بسیار زیاد جهت برچسب‌گذاری دستی پایگاه داده، شبکه دانش‌آموز عملکرد بسیار نزدیکی به نسبت مدل آموزش داده شده توسط انسان از خود نشان می‌دهد. اختلاف عملکرد مدل دانش‌آموز در کلاس‌های وسیعی نظیر

جاده، آسمان و درخت بـه‌ترتیـب ۳/۸٪، ۵/۱٪ و ۳/۳٪ بـه‌دسـت آمده است. همچنین در کلاس‌های چالشی‌تر نظیر عـابرین پیـاده که ممکن است شکل‌ها و کانتورهای بسیار گسترده‌ای را به خـود اختصاص دهند اختلاف عملکرد شبکه دانـش‌آمـوز ۳٪ محاسـبه گشته است. در ارزیابی کلی نیز مدل FastSCNN هنگام آمـوزش دیدن از DABNet و بدون هیچگونه دخالتی در برچسـب‌گـذاری تصاویر آموزش خود، تنها اخـتلاف عملکـرد ۴/۵٪ در مقایسـه بـا مدل متناظری داشته است که با صرف زمان بسیار توسط انسـان آموزش داده شده است.

۶- مراجع

[1] S. Singh, “Critical reasons for crashes investigated in the national motor vehicle crash causation survey,” Traffic Saf. Facts - Crash Stats, 2015.

[2] F. Becker and K. W. Axhausen, “Literature review on surveys investigating the acceptance of automated vehicles,” in TRB 96th Annual Meeting Compendium of Papers, pp. 1–12 , 2017

[3] C. Gkartzonikas and K. Gkritza, “What have we learned? A review of stated preference and choice studies on autonomous vehicles,” Transp. Res. Part C, vol. 98, pp. 323–337, 2019.

[4] J. Cui, L. S. Liew, et al. “A review on safety failures, security attacks, and available counter measures for autonomous vehicles,” Ad. Hoc. Networks, vol. 90, p. 101823, 2019.

[5] J. Van Brummelen, M. O’Brien, et al. “Autonomous vehicle perception: The technology of today and tomorrow,” Transp. Res. Emerg. Technol., Part C, vol. 89, pp. 384–406, 2018.

[6] J. Janai, F. Guney, et al. “Computer vision for autonomous vehicles: Problems, datasets and state of the art,” Foundations and Trends in Computer Graphics and Vision, vol. 12, no. 1-3, pp. 1-308, 2020.

[7] D. Feng, et al. “Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges,” IEEE Transactions on Intelligent Transportation Systems, 2020, DOI: 10.1109/TITS.2020.2972974.

[8] K. Kim, J. S. Kims S. Jeong, et al. “Cybersecurity for autonomous vehicles: Review of attacks and defense,” Computers & Security, vol. 103, p. 102150, 2021.

[9] Z. El Rewini, K. Sadatsharan, D. F. Selvaraj, et al., “Cybersecurity challenges in vehicular communications,” Vehicular Communications, vol. 23, p. 100214, 2020.

[10] C. Kamann and C. Rother, “Benchmarking the robustness of semantic segmentation models,” in arXiv preprint arXiv:1908.05005, 2019.

[11] H. Wu, Y. Yan, Y. Ye, M. K. Ng, and Q. Wu, “Geometric knowledge embedding for unsupervised domain adaptation,” Knowledge-Based Systems, vol. 191, p. 105155, 2020.

[12] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3339-3348.

[13] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in IEEE Intelligent Vehicles Symposium (IV), 2019, pp. 1312-1318.

[14] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213-3223.

[15] F. Yu et al., “BDD100K: a diverse driving dataset for heterogeneous multitask learning,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2636-2645.

[16] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361.

[17] D. Heo, J. Nam, and B. Ko, “Estimation of pedestrian pose orientation using soft target training based on teacher–student framework,” Sensors, vol. 19, no. 5, p. 1147, 2019.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.

[19] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801-818.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition.” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[21] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W. S. Zheng, “Improving fast segmentation with teacher-student learning,” in British Machine Vision Conference (BMVC), 2018, pp. 205.

[22] D. Heo, J. Nam, and B. Ko, "Estimation of pedestrian pose orientation using soft target training based on teacher–student framework," Sensors, vol. 19, no. 5, pp. 1147, 2019.

[23] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: common practices and emerging technologies," IEEE Access, vol. 8, pp. 58443-58469, 2020.

[24] Y. Zhu et al., "Improving semantic segmentation via self-training," arXiv preprint arXiv: 2004.14960, 2020.

[25] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in Proceedings of European Conference on Computer Vision (ECCV), 2008, pp. 44-57.

[26] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classesin video: A high-definition ground truth database," Pattern Recognition, vol. 30, no. 2, pp. 88–97, 2009.

[27] L.-C. Chen et al., "Naive-student: leveraging semi-supervised learning in video sequences for urban scene segmentation," in Proceedings of European Conference on Computer Vision (ECCV), 2020, pp. 695-714.

[28] Q. Xie, M.-T. Luong, E. Hovy, and Q. V Le, "Self-training with noisy student improves imageNet classification," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10687-10698.

[29] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3234-3243.

[30] Y. H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1456-1465.

[31] Y. H. Tsai, W. C. Hung, S. Schulter, K. Sohn, M. H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7472-7481.

[32] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2090-2099.

[33] T. H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "Advent: adversarial entropy minimization for domain adaptation in semantic segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2517-2526.

[34] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3764-3773.

[35] U. Michieli, M. Biasetton, G. Agresti, and P. Zanuttigh, "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," IEEE Transactions on Intelligent Vehicles, , vol. 5, no. 3, pp. 508-518, 2020.

[36] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: ground truth from computer games," in Proceedings of European Conference on Computer Vision (ECCV), 2016, pp. 102-118.

[37] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 4990-4999.

[38] S. M. Khorashadizadeh, V. Azadzadeh, and, A. M. Latif, "Detection of pornographic digital images using support vector machine and neural network," Journal of Electronical & Cyber Defence, vol. 4, no. 4, pp. 79-88, 2017. (in Persian)

[39] M. Asadi, M. A. Jabraeil Jamali, et al., "Comparison of supervised machine learning algorithms in detection of botnets domain generation algorithms," Journal of Electronical & Cyber Defence, vol. 8, no. 4, pp. 17-29, 2020. (in Persian)

[40] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation," in British Machine Vision Conference (BMVC), 2019.

[41] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: exploring context and detail for semantic segmentation in real-time," in British Machine Vision Conference (BMVC), 2018.

[42] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: fast semantic segmentation network," in British Machine Vision Conference (BMVC), 2019.

[43] D. Mazzini, "guided upsampling network for real-time semantic segmentation," in British Machine Vision Conference (BMVC), 2018, p. 117.

[44] C. Yu, "BiSeNet: bilateral segmentation network for real-time semantic segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 325-341.

[45] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation mark," arXiv preprint arXiv:1801.04381, 2018.

[48] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in Proceedings of European Conference on Computer Vision (ECCV), 2018, pp. 405-420.

[49] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: generalization gap and sharp minima," Int. Conf. Learn. Represent, pp. 1–16. 2016

[50] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations (ICLR), 2015.

[51] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: an open urban driving simulator," in Conference on Robot Learning (CoRL), 2017.

[52] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 2017.

[53] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[54] M. Liu and H. Yin, "Feature pyramid encoding network for real-time semantic segmentation," arXiv preprint arXiv:1909.08599, 2019.

[55] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 552-568.

[56] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9190-9200.