نشریه علمی «علوم و فناوری‌های پدافند نوین»
سال یازدهم، شماره ۱، بهار ۱۳۹۹؛ ص ۵۹-۵۱

تخمین جریان نوری با استفاده از تقسیم‌بندی معنایی و شبکه عصبی عمیق AES

هانیه زمانیان[۱]، حسن فرسی[۲*]، سجاد محمدزاده[۳]

۱- دانشجوی دکتری، ۲- استاد، ۳- استادیار، دانشگاه بیرجند

(دریافت: ۹۸/۰۲/۲۴، پذیرش: ۹۸/۰۵/۰۸)

چکیده

اهمیت و نیاز به درک صحنه‌های بصری به علت پیشرفت سامانه‌های خودکار به‌طور پیوسته افزایش یافته است. جریان نوری یکی از ابزارهای درک صحنه‌های بصری است. روش‌های جریان نوری موجود، مفروضات کلی و همگن فضایی، در مورد ساختار فضایی جریان نوری ارائه می‌دهند. درواقع، جریان نوری در یک تصویر، بسته به کلاس شی و همچنین نوع حرکت اشیاء مختلف، متفاوت است. فرض اول در میان بسیاری از روش‌ها در این زمینه، پایداری روشنایی در طی حرکت پیکسل‌ها بین فریم‌ها است. ثابت شده است که این فرض در حالت کلی صحیح ناست. در این پژوهش از تقسیم‌بندی اشیای موجود در تصویر و تعیین حرکت اشیا به‌جای حرکت پیکسلی کمک گرفته شده است. درواقع از پیشرفت‌های اخیر شبکه‌های عصبی کانولوشن در تقسیم‌بندی معنایی صحنه‌های استاتیک، برای تقسیم تصویر به اشیا مختلف بهره گرفته می‌شود و الگوهای مختلف حرکتی بسته به نوعشی تعریف می‌شود. سپس، تخمین جریان نوری با استفاده از ایجاد یک شبکه عصبی کانولوشن عمیق برای تصویری که در مرحله اول تقسیم‌بندی معنایی شده است، انجام می‌شود. روش پیشنهادی کمترین خطا در معیار جریان نوری برای پایگاه داده KITTI-2015 را فراهم می‌آورد و تقسیم‌بندی بهتری را نسبت به روش‌های اخیر در طیف وسیعی از فیلم‌های طبیعی تولید می‌کند.

کلیدواژه‌ها: جریان نوری، تقسیم‌بندی معنایی، شبکه عصبی عمیق، رمزگذار، رمزگشا

Estimation of Optical Flow using Semantic Segmentation and AES Deep Neural Network

H. Zamanian, H. Farsi[*], S. Mohamadzadeh

University of Birjand

(Received: 14/05/2019; Accepted: 30/07/2019)

Abstract

The importance and demand of visual scene understanding have been increasing because of autonomous systems development. Optical flow is known as an important tool for scene understanding. Current optical flow methods present general assumptions and spatial homogeneous for spatial structure of flow. In fact, the optical flow in an image depends on object class and the type of object movement. The first assumption in many methods in this field is the brightness constancy during movements of pixels between frames. This assumption is proven to be inaccurate in general. In this paper, we use recent development of deep convolutional networks in semantic segmentation of static scenes to divide an image in to different objects and also depends on type of the object different movement patterns are defined. Next, estimation of the optical flow is performed by using deep neural network for initial image which has been semantically segmented. The proposed method provides minimum error in optical flow measures for KITTI-2015 database and results in more accurate segmentation compared to state-of-the-art methods for several natural videos.

Keywords: Optical Flow, Semantic Segmentation, Deep Neural Network, Encoder, Decoder

* Corresponding Author E-mail: hfarsi@birjand.ac.ir

۱. مقدمه

هدف جریان نوری[1]، تخمین یک بردار دوبعدی، به‌منظور کد کردن حرکت بین دو فریم متوالی برای هر پیکسل است. معمولاً فرض می‌شود که یک منطقه محلی در اطراف هر پیکسل در هر دو فریم متوالی مشابه یکدیگر است. علیرغم تحقیقات انجام‌شده در دهه اخیر، تخمین جریان نوری متراکم به‌عنوان یک مشکل بزرگ هنوز شناخته می‌شود. جابجایی‌های زیاد، مناطق بدون بافت، ناهنجاری‌ها، سایه‌ها و تغییرات شدید در روشنایی ازجمله عوامل ایجاد این مشکل هستند [۱]. علاوه براین، ازآنجایی‌که ممکن است بیش از ۳۰ هزار احتمال برای پتانسیل حرکت پیکسل در نظر گرفته شود، تخمین جریان نوری نیازمند محاسبات زیاد است. این مسئله برای روش‌های گسسته مشکلاتی را ایجاد می‌کند، بنابراین، بیشتر روش‌های اخیر از روش‌های بهینه‌سازی پیوسته استفاده می‌کنند [۲].

نتایج حاصل از چندین پایگاه داده اخیر نشان می‌دهد که دقت روش‌های جریان نوری به‌طور پیوسته بهبود یافته است [۳]. بااین‌حال، حتی روش‌های جریان نوری پیشرفته در مورد تصاویر با حرکات سریع، در مناطقی با بافت کم و در اطراف مرزهای جسم (انسداد) ضعیف عمل می‌کنند. هدف این پژوهش، بهبود تخمین جریان نوری با استفاده از تقسیم‌بندی معنایی تصویر است. با توجه به تحقیقات انجام‌شده، پیشرفت‌های زیادی در زمینه تقسیم‌بندی معنایی صورت گرفته است که ازجمله تخمین جریان نوری با استفاده از شبکه‌های عصبی کانولوشن[2] (CNN) است که حجم زیادی از داده‌های برچسب‌گذاری شده را شامل می‌شود. در این پژوهش یک شبکه CNN جدید برای تقسیم‌بندی معنایی تصویر که نتایج بهتری نسبت به سایر روش‌های موجود ارائه داده است، معرفی می‌شود و سپس از این شبکه برای بهبود تخمین جریان نوری کمک گرفته می‌شود [۴].

تقسیم‌بندی معنایی تصویر به چند دلیل می‌تواند موجب بهبود تخمین جریان نوری شود. ۱) اطلاعات مربوط به مرزهای شی را فراهم می‌کند، ۲) ازآنجایی‌که اشیا مختلف به‌طور متفاوت حرکت می‌کنند؛ به‌عنوان مثال، جاده‌های مسطح بی‌حرکت هستند، اتومبیل‌ها به‌طور مستقل حرکت می‌کنند و درختان در باد نوسان می‌کنند، باید تخمین حرکت و درنتیجه جریان نوری بین مناطق با برچسب‌های کلاسی مختلف، متفاوت باشد [۵].

در این مقاله، هدف محاسبه جریان نوری در زمینه رانندگی مستقل است. در این شرایط خاص، صحنه‌ها اغلب از یک پس‌زمینه استاتیک و تعداد کمی اجزا متحرک در ترافیک تشکیل شده است. برای بهره‌برداری از چنین وضعیتی، باید اشیا متحرک و مستقل از هم شناسایی شوند و حرکت آن‌ها تخمین زده شود. روش‌های گذشته، به‌طورمعمول تلاش می‌کنند که اشیا را تنها برحسب نوع حرکت دسته‌بندی کنند اما همان‌طور که می‌دانیم تخمین حرکت برای تقسیم‌بندی دقیق حرکت ضروری است.

در مقابل، در این مقاله یک رویکرد جایگزین پیشنهاد می‌شود که تنها با بهره‌برداری از روش‌های شناخت نوع اشیا، برای شناسایی اشیاء بالقوه متحرک به نتایج مطلوبی دست پیدا کند. باید توجه داشت که تقسیم‌بندی معنایی کافی نیست، زیرا وسایل مختلف ممکن است بسیار متفاوت حرکت کنند اما به دلیل انسداد، یک مؤلفه متصل را تشکیل دهند. بنابراین، در اینجا، از تقسیم‌بندی معنایی به‌گونه‌ای استفاده شده است که درنتیجه آن برای هر خودرو یک برچسب تقسیم‌بندی متفاوت در نظر گرفته می‌شود که به آن تقسیم‌بندی نمونه‌ای گفته می‌شود.

با توجه به تقسیم‌بندی نمونه‌ای، مشکل تخمین جریان نوری برای هر جسم متحرک، به مجموعه‌ای از مسائل تخمین حرکت دوبعدی تبدیل می‌شود. پس‌زمینه به‌عنوان یک شی خاص محسوب می‌شود که حرکات آن صرفاً به خاطر شخص ثالث است. این رویکرد متفاوت با روش‌های تخمین جریان موجود است، که در آن‌ها صحنه استاتیک فرض می‌شود و تنها مشاهده‌کننده حرکت می‌کند [۵ و ۶]. همان‌طور که در بخش ارزیابی تجربی نشان داده‌شده است، نتایج روش پیشنهادی، تخمین جریان بسیار بهتری برای حرکت اشیا نشان می‌دهد.

در این پژوهش از دو شبکه عصبی کانولوشن، که به‌صورت متوالی قرار گرفته‌اند استفاده می‌شود که شبکه اول مربوط به تقسیم‌بندی معنایی تصویر است و خروجی این شبکه به‌عنوان ورودی به شبکه بعدی که وظیفه تخمین جریان نوری را بر عهده دارد، داده می‌شود. بنابراین، در این تحقیق، دو کار مهم انجام می‌شود، ۱) روش تخمین جریان نوری ارائه و پیشنهاد می‌شود که از این اطلاعات معنایی در مورد صحنه‌ها، اشیا و تقسیم‌بندی آن‌ها استفاده می‌شود که تولید کمترین خطا را در مقایسه با روش‌های تک سویی بر پایه معیار جریان KITTI 2015 ارائه می‌دهد [۴]. ۲) نتایج این مقاله نشان داده است که دانستن نوع و محل اشیا به تخمین حرکت آن‌ها کمک می‌کند.

رویکرد کلاسیک برای تخمین جریان نوری شامل ایجاد یک مدل انرژی می‌شود، که معمولاً شواهد تصویری مانند قاعده گرادیان، انحراف و یا تطابق را ترکیب می‌کند [۱ و ۷-۹]. سان و همکاران [۲] در تحقیقشان نشان می‌دهند که روش‌های کلاسیک برای تخمین جریان نوری عمدتاً روش‌های مبتنی بر گرادیان هستند. متأسفانه، این روش‌ها به‌طورمعمول برای تخمین

[1] Optical Flow
[2] Convolutional Neural Network (CNN)

جابجایی‌های بزرگ (که اغلب در صحنه‌های ترافیکی دیده می‌شود) به دلیل ظاهر متناقض تکه‌های تصویر، نامناسب هستند [۷ و ۸]. یک روش دیگر EpicFlow است که اغلب برای تخمین جریان‌های ضعیف با توجه به لبه‌ها استفاده می‌شود [۱ و ۴].

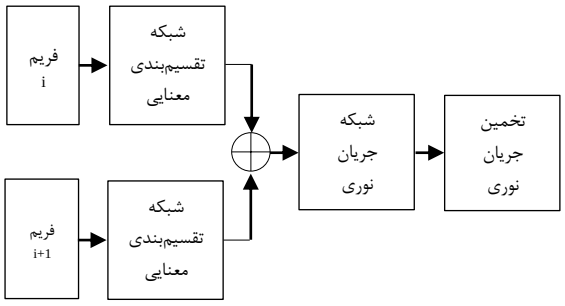
بسیاری از رویکردهای تخمین جریان را استنتاج در میدان تصادفی مارکف[1] (MRF) تشکیل می‌دهند [۱۲-۱۰]. معمولاً برای استنتاج، از الگوریتم‌های ارسال پیام یا ایجاد حرکت استفاده می‌شود. یکی از موفق‌ترین روش‌های جریان نوری در زمینه رانندگی مستقل، روش DiscreteFlow است که فضای جستجو را تنها با استفاده از تعداد کمی از پیشنهادها کاهش می‌دهد [۱۲]. سپس از MRF برای یافتن نتایج صحیح استفاده می‌شود. پس از اعمال کمی پردازش ، جریان نوری نهایی با استفاده از EpicFlow محاسبه و نمایش داده می‌شود. یان و لی [۱۳] در مطالعات خود تصاویر را با استفاده از سوپر پیکسل‌ها تقسیم‌بندی می‌کنند و جریان هر یک از ابر پیکسل‌ها را به‌عنوان همگرایی‌های صفحات سه‌بعدی تخمین می‌زنند. این روش‌ها از این واقعیت که پس‌زمینه استاتیک است و تنها چند شی حرکت می‌کنند، بهره نمی‌برند. در روش پیشنهادی در این مقاله سعی شده است که از این موضوع استفاده شود.

تاکنون پژوهش‌های زیادی برای تخمین جریان نوری و تقسیم‌بندی معنایی به‌طور هم‌زمان صورت گرفته است. روش‌های متداول‌تر، از تقسیم‌بندی تصویر برای کمک به جریان نوری استفاده می‌کنند. یان و لی [۱۳] صحنه را با توجه به رنگ یا سایر نشانه‌ها به قسمت‌هایی تقسیم می‌کنند و سپس مدل‌های جریان پارامتریک را در آن قرار می‌دهند. نوع مدل در هر منطقه متفاوت است، اما در این پژوهش گامی فراتر برداشته تا از اطلاعات معنایی برای تعیین مدل مناسب استفاده شود. سان و همکارانش [۱۴] ابتدا صحنه را به ابرپیکسل‌ها تقسیم می‌کنند و سپس در مورد روابط انسدادی بین ابرپیکسل‌هایی که در همسایگی یکدیگر هستند بحث می‌کنند. این روش‌ها هیچ اطلاعاتی در مورد اشیا تقسیم‌بندی شده ندارند، بلکه به دنبال تقسیم‌بندی صحنه به مناطق متحرک هستند. سویلا و همکاران [۱۵] نشان داده‌اند که از تقسیم‌بندی معنایی نیز می‌توان برای کمک به جریان نوری استفاده کرد. به‌طور خاص، با تقسیم‌بندی معنایی، اشیا به سه طبقه دسته‌بندی می‌شوند: پس‌زمینه استاتیک مسطح، اشیا متحرک و عناصری که مدل حرکت خاصی برای آن‌ها نمی‌توان تعریف کرد. سپس برای هر یک از سه کلاس یک مدل جریان متفاوت با استفاده از DiscreteFlow اقتباس می‌شود.

[1] Markov Random Field

۲. روش تحقیق

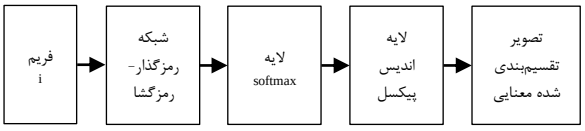
پژوهش‌های انجام‌گرفته تاکنون نشان داده‌اند که شبکه‌های CNN می‌توانند در کارهایی مانند طبقه‌بندی، تقسیم‌بندی معنایی و تشخیص اشیا بسیار خوب عمل کنند [۱۶]. اخیراً از این شبکه‌ها برای تطبیق استریو استفاده شده است که عملکرد مناسبی را به همراه داشته‌اند [۱۷ و ۱۸] و نتایج به‌دست‌آمده از سایر روش‌ها برای پایگاه داده KITTI را به چالش کشیده‌اند [۱۹]. به دنبال این روند، در این پژوهش نیز با طراحی یک شبکه عصبی پیچیده عمیق، ویژگی‌هایی که برای تخمین جریان نوری مطلوب‌اند استخراج می‌شوند. شبکه پیشنهادی دارای دو بخش عمده است. بخش اول وظیفه تقسیم‌بندی معنایی تصویر را به عهده دارد و نتایج این بخش به‌عنوان ورودی بخش دوم در نظر گرفته می‌شود. بخش دوم نیز وظیفه تخمین جریان نوری با توجه به اطلاعات بخش اول را به عهده دارد. شکل (۱) نمای کلی شبکه پیشنهادی را برای دو فریم متوالی i و i+1 نشان می‌دهد.



شکل ۱. بلوک دیاگرام شبکه پیشنهادی

۲-۱. تقسیم‌بندی معنایی

در این پژوهش از نرم‌افزار MATLAB 2017 برای آموزش و آزمایش یک شبکه جدید CNN برای تقسیم‌بندی معنایی استفاده شده است. یکی از نوآوری‌های این پژوهش استفاده از این شبکه است که بلوک دیاگرام آن در شکل (۲) نمایش داده شده است. این شبکه از لایه‌های رمزگذار[2] و رمزگشا[3]، لایه Softmax و لایه اندیس پیکسل[4] تشکیل شده است که ایده کلی آن از شبکه متداول SegNet گرفته شده است [۲۰].



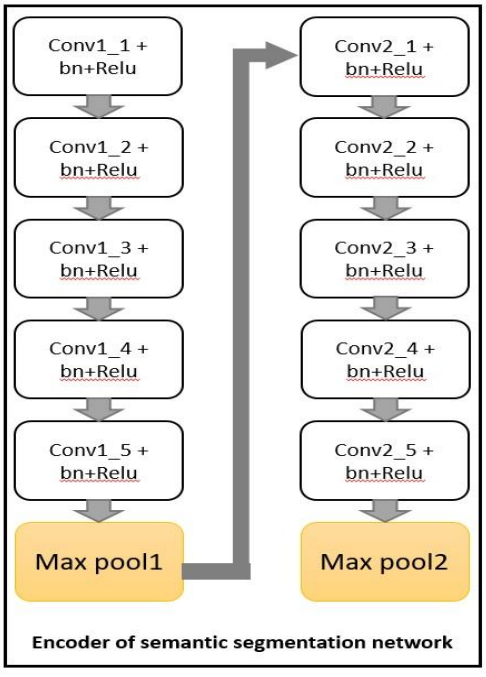
شکل ۲. بلوک دیاگرام شبکه تقسیم‌بندی معنایی

با توجه به قابلیت‌هایی که تاکنون الگوریتم SegNet در بهبود

[2] Encoder
[3] Decoder
[4] Pixel lable

نتایج تقسیم‌بندی معنایی از خود نشان داده، رویکرد این مقاله بر روی تنظیم بهتر پارامترها و تعداد لایه‌های این الگوریتم است. روش پیشنهادی، یک روش رمزگذار-رمزگشا است که از الگوریتم Segnet الهام گرفته و برای تقسیم‌بندی معنایی طراحی شده است. از دیدگاه محاسباتی، شبکه طراحی‌شده باید از نظر حافظه موردنیاز و مدت‌زمان انجام محاسبات در مرحله استنتاج کارآمد باشد. مسلماً هرچه تعداد لایه‌ها و پارامترهای قابل یادگیری کم‌تر باشد، شبکه از دیدگاه محاسباتی کارآمدتر است و به‌علاوه برای کاربردهای برخط مفیدتر خواهد بود. هدف الگوریتم ارائه‌شده افزایش دقت در تقسیم‌بندی معنایی همراه با کاهش این پارامترها است.

در روش پیشنهادی، بخش رمزگذار دارای عمق دو است، که در اولین بخش از ۵ لایه کانولوشن که هرکدام دارای ۶۴ فیلتر با ابعاد ۳×۳ است، استفاده شده است. بعد از هر لایه کانولوشن، یک لایه نرمال‌سازی batch (BN)[1] و سپس یک لایه ReLU[2] قرار دارد. در بخش دوم رمزگذاری نیز عیناً مراحل رمزگذار اول اجرا می‌شود و در ادامه آن بخش رمزگشای است که ابعاد فیلترهای رمزگشا، متناسب با کانولوشن‌های اجراشده در هر مرحله از رمزگذاری تنظیم می‌شود. یعنی در هر گام از رمزگشایی از ۶۴ فیلتر با ابعاد ۳×۳ استفاده می‌شود که وزن‌های این فیلترها باید با آموزش شبکه و متناسب با داده‌های آموزشی تنظیم شوند. در شکل (۳)، معماری رمزگذاری شبکه پیشنهادی نمایش داده شده است.

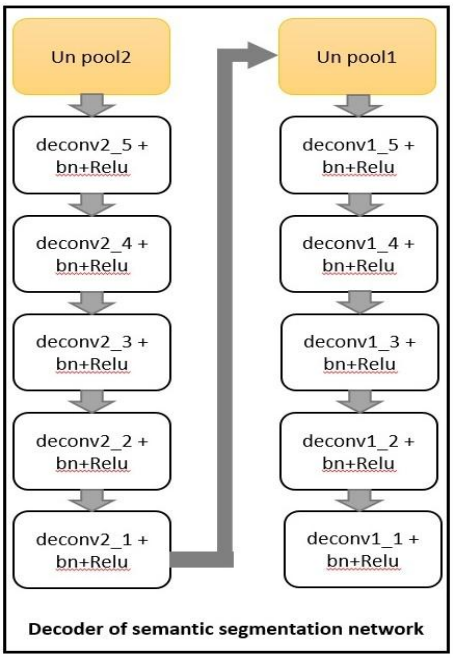


شکل ۳. معماری رمزگذاری شبکه تقسیم‌بندی معنایی

[1] Batch Normalization
[2] Rectified Linear Unit

سپس عمل رمزگشایی که عکس عمل رمزگذاری است، انجام می‌شود، تا تصویر نهایی همراه با تقسیم‌بندی معنایی اشیا تصویر بر اساس ویژگی‌های استخراجی شبکه بازیابی شود. در شکل (۴)، معماری بخش رمزگشایی شبکه پیشنهادی نمایش داده شده است.

در انتهای رمزگشایی مرحله دوم، خروجی با ابعادی برابر با تصویر ورودی ایجاد می‌شود که در لایه softmax پاسخ نهایی تقسیم‌بندی مشخص می‌شود و در واقع عمل طبقه‌بندی پیکسلی انجام می‌شود. با توجه به بزرگ بودن ابعاد تصویر ورودی و درنتیجه تعداد زیاد پیکسل‌هایی که باید در مرحله طبقه‌بندی در مورد آن‌ها تصمیم‌گیری شود، بیش‌ترین پارامتر قابل تنظیم، در این مرحله وجود دارد.

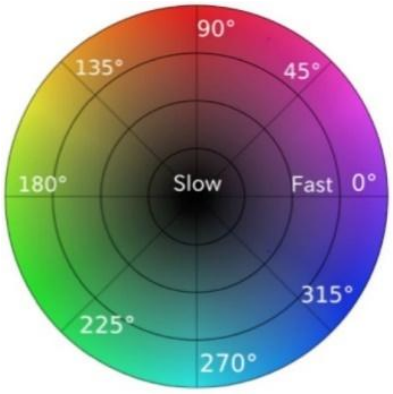


شکل ۴. معماری رمزگشایی شبکه تقسیم‌بندی معنایی

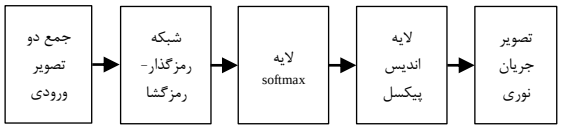
۲-۲. جریان نوری

رویکرد این پژوهش پردازش جداگانه اما یکسان برای دو تصویری است که موردمحاسبه جریان نوری هستند و سپس ویژگی‌های استخراجی از آن‌ها در مرحله بعدی ترکیب می‌شوند. با استفاده از این معماری، ابتدا ویژگی‌های معنی‌دار دو عکس به‌طور جداگانه تولید می‌شوند و سپس آن ویژگی‌ها در یک سطح بالاتر ترکیب می‌شوند. در این روش ابتدا دو تصویری که در طی مرحله قبل تقسیم‌بندی معنایی شده‌اند به شبکه تخمین جریان نوری داده می‌شوند. در مرحله اول دو تصویر متوالی با یکدیگر جمع شده و سپس وارد شبکه CNN تخمین جریان نوری می‌شوند. از این شبکه برای تخمین ۱۶ جهت جریان استفاده شده است. در واقع

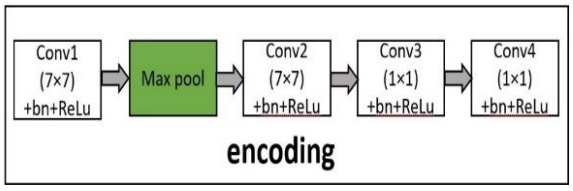
برای تخمین جریان نوری، مانند تقسیم‌بندی معنایی از طبقه‌بندی استفاده شده است. بنابراین، این شبکه نیز همانند شبکه پیشنهادی برای تقسیم‌بندی معنایی از بخش‌های رمزگذار و رمزگشا تشکیل شده است. در شکل (۵) تصویری از نحوه تقسیم‌بندی رنگ‌های جریان نوری برای ایجاد کلاس‌های مورداستفاده در طبقه‌بندی نشان داده شده است [۲۱]. همان‌طور که مشاهده می‌شود سرعت و جهت حرکت، دو پارامتر مؤثر در طبقه‌بندی جریان نوری هستند. شکل (۶) نیز نمایشی از شبکه پیشنهادی برای تخمین جریان نوری را ارائه می‌دهد. جزییات بخش‌های رمزگذاری و رمزگشایی نیز در شکل‌های (۷) و (۸) ارائه شده است.



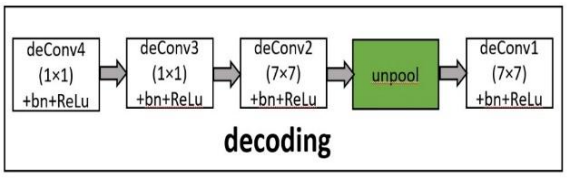
شکل ۵. نحوه تقسیم‌بندی رنگ‌های جریان نوری برای ایجاد کلاس‌های مورداستفاده در طبقه‌بندی بر اساس سرعت و جهت حرکت [۲۱]



شکل ۶. بلوک دیاگرام شبکه تخمین جریان نوری



شکل ۷. معماری رمزگذاری شبکه تخمین جریان نوری



شکل ۸. معماری رمزگشایی شبکه تخمین جریان نوری

در مرحله رمزگذاری از ۴ لایه کانولوشن که هرکدام دارای ۳۲ فیلتر هستند و ابعاد فیلترهای هر لایه در بلوک مربوط به آن لایه همان‌طوری که در شکل (۷) نشان داده شده است، استفاده شده است. بعد از هر لایه کانولوشن یک لایه BN و سپس یک لایه ReLU قرار دارد. سپس عمل رمزگشایی که عکس عمل رمزگذاری است، انجام می‌شود، تا تصویر نهایی همراه با طبقه‌بندی تصویر مطابق با جهت و سرعت جریان نوری، که در بخش قبل توضیح داده شد، بر اساس ویژگی‌های استخراجی شبکه بازیابی شود. در انتهای مرحله رمزگشایی، خروجی با ابعادی برابر با تصویر ورودی ایجاد می‌شود که در لایه softmax پاسخ نهایی طبقه‌بندی مشخص می‌شود.

۲-۳. پایگاه داده

در این پژوهش، از پایگاه داده‌های صحنه جاده‌ای CamVid [۲۲] برای ارزیابی عملکرد انواع روش‌ها در بخش تقسیم‌بندی معنایی استفاده شده است. این پایگاه داده شامل ۷۰۱ تصویر است که از میان آن‌ها ۴۲۱ تصویر برای آموزش و ۲۸۰ تصویر برای آزمایش و اعتبارسنجی استفاده شده است. این تصاویر رنگی هستند و شامل صحنه‌های روز و غروب، با رزولوشن ۳۶۰×۴۸۰ پیکسل هستند [۲۲]. چالش این پایگاه داده این است که ۱۱ کلاس که شامل جاده، ساختمان، اتومبیل، عابر پیاده، علائم، ستون‌ها، پیاده‌رو و دوچرخه‌سوار، آسمان، درخت و فنس است، از یکدیگر جدا شوند [۲۲].

در مورد شبکه کامل که شامل هر دو بخش تقسیم‌بندی معنایی و تخمین جریان نوری است از پایگاه داده معروف KITTI 2015 استفاده شده است [۴]. این پایگاه داده شامل تصویر مرجع برای هر دو بخش تقسیم‌بندی معنایی و جریان نوری است که مزیت اصلی این پایگاه داده برای ایده ارائه‌شده در این پژوهش است. به‌علاوه، چون این پایگاه داده نیز شامل صحنه‌های جاده‌ای و ترافیکی است، کافی است که هنگام آموزش شبکه از وزن‌های به‌دست‌آمده در زمان آموزش با پایگاه داده CamVid به‌عنوان وزن اولیه استفاده شود، که در این صورت سرعت آموزش بسیار افزایش می‌یابد.پایگاه داده KITTI 2015 شامل ۴۰۰ تصویر است که ۲۰۰ تصویر برای آموزش و ۲۰۰ تصویر برای آزمایش در مرحله تقسیم‌بندی معنایی استفاده شده است و در مرحله تخمین جریان نوری ۴۰۰ جفت فریم است که هر جفت شامل دو فریم متوالی است و مجدداً ۲۰۰ جفت فریم برای آموزش و ۲۰۰ جفت فریم برای آزمایش در نظر گرفته است. ابعاد تصاویر این پایگاه داده ۳۷۶×۱۲۴۱ پیکسل است.

۲-۴. معیارهای مقایسه

برای مقایسه عملکرد کمی انواع روش‌ها در مرحله تقسیم‌بندی معنایی، از سه معیار زیر استفاده شده است.

۱) دقت کلی GA[1]، که درصد پیکسل‌های صحیح طبقه‌بندی‌شده در پایگاه داده را اندازه‌گیری می‌کند.

$$GA = \frac{P_C}{N} \times 100\% \quad (۱)$$

که در آن P_c تعداد پیکسل‌های صحیح طبقه‌بندی‌شده و N تعداد کل پیکسل‌های تصویر است [۲۳].

۲) دقت کلاسی CA[2]، که میانگین دقت پیش‌بینی برای تمام کلاس‌ها را محاسبه می‌کند.

$$CA = \frac{1}{M}\sum_{i=1}^{M}\frac{P_{C_i}}{P_{t_i}} \times 100\% \quad (۲)$$

که در آن P_{c_i} تعداد پیکسل‌های صحیح طبقه‌بندی‌شده در کلاس i-ام و P_{t_i} تعداد کل پیکسل‌های کلاس i-ام تصویر با M کلاس مجزا است [۲۳]. در واقع این معیار نوعی میانگین‌گیری از دقت طبقه‌بندی در کلاس‌های مختلف است.

۳) میانگین تقاطع کلی[3] (mIoU) در تمام کلاس‌ها، که در چالش Pascal VOC12 مورداستفاده قرار گرفته است [۲۴]. اگر A_i ناحیه تقسیم‌بندی شده در تصویر مرجع برای کلاس i-ام و B_i ناحیه تقسیم‌بندی شده توسط الگوریتم مورداستفاده برای کلاس i-ام باشد، میانگین تقاطع کلی، طبق فرمول ۳، برای M کلاس محاسبه می‌گردد.

$$mIoU = \frac{1}{M}\sum_{i=1}^{M}\frac{A_i \bigcap B_i}{A_i \bigcup B_i} \quad (۳)$$

معیار mIoU دقیق‌تر از دقت متوسط کلاس است، زیرا پیش‌بینی‌های مثبت کاذب را جریمه می‌کند [۲۳].

۴) برای ارزیابی مرحله تخمین جریان نوری از معیار خطای نقطه انتهایی جریان (AEE)[4] استفاده شده است که به شکل زیر محاسبه می‌شود [۲۵].

$$AEE = \sqrt{(u - u_{GT})^2 + (v - v_{GT})^2} \quad (۴)$$

که در آن، u و v مقادیر سرعت و جهت حرکت در تخمین جریان نوری توسط شبکه پیشنهادی هستند. در واقع این مقادیر به ترتیب نشان‌دهنده محورهای افقی و عمودی در شکل (۵) بوده و u_{GT} و v_{GT} مقادیر سرعت و جهت حرکت در تصویر مرجع در یک پایگاه داده دلخواه هستند [۲۵].

۵) معیار دیگری که برای بررسی تخمین جریان نوری ارائه شده است، FL-all است. اگر تفاوت مقدار تخمین زده‌شده برای جریان نوری با تصویر مرجع جریان نوری بیش از ۵ درصد و یا خطای نقطه پایانی جریان بیش از ۳ پیکسل باشد، آن پیکسل به‌عنوان خطا در نظر گرفته می‌شود. نسبت تعداد پیکسل‌های خطا به تعداد کل پیکسل‌ها برحسب درصد، معیار FL-all را نشان می‌دهد [۲۶].

[1]Global Accuracy
[2]Class Accuracy
[3] Mean Intersection Over Union (mIoU)
[4]Average Endpoint Error

۲-۵. آموزش شبکه

برای آموزش این شبکه نیز مشابه شبکه پیشنهادی برای تقسیم‌بندی معنایی، از کاهش گرادیان تصادفی[5] (SGD) با نرخ یادگیری اولیه ۰/۱ و کاهش آن با ضریب ۰/۱ بعد از هر ۱۰۰ دوره (۲۰ هزار تکرار) و حرکت[6] ۰/۵، استفاده شده است. بعلاوه، از تلفات آنتروپی متقاطع، به‌عنوان تابع هدف برای آموزش شبکه استفاده شده است [۲۷].

برای آموزش و آزمون شبکه پیشنهادی، از پردازنده CPU با مشخصات Intel Core i7- 6700HQ و کارت گرافیک NVIDIA GEFORCE GTX 950M و یک GPU و همچنین ۱۲ گیگابایت حافظه استفاده شده است. کدهای روش پیشنهادی با استفاده از نرم‌افزار MATLAB نوشته شده است.

۳. نتایج و بحث

همان‌طور که قبلاً نیز اشاره شد، شبکه پیشنهادی در بخش تقسیم‌بندی معنایی با استفاده از تصاویر پایگاه داده‌های CamVid مورد آموزش و آزمایش قرار گرفته است. این شبکه به علت تعداد پارامترهای کمی که نسبت به سایر الگوریتم‌های مشابه دارد، دارای سرعت یادگیری بسیار بیشتری است. درعین‌حال که توانسته است عملکرد بهتری ازلحاظ دقت تقسیم‌بندی نیز از خود نشان دهد، که نتایج به‌دست‌آمده بیانگر این ادعا است. تعداد پارامترهای قابل یادگیری در این شبکه در مقایسه با الگوریتم متداول SegNet که دارای ۱۴/۷ مگا پارامتر است، تنها ۱/۴۱ مگا پارامتر خواهد بود.

نحوه محاسبه تعداد پارامترهای قابل یادگیری در لایه‌های یک شبکه CNN بدین صورت است که هر لایه کانولوشنی با تعداد l نقشه ویژگی در ورودی و k نقشه ویژگی در خروجی و ابعاد فیلتر n×m، دارای (n×m×l+1)×k پارامتر قابل یادگیری است. لایه‌های pooling پارامتر قابل یادگیری ندارند. لایه‌های تماماً متصل مانند softmax و لایه خروجی، با تعداد n ورودی و m خروجی دارای (n+1)×m پارامتر قابل یادگیری است.

جدول (۱)، مقایسه‌ای بین تعداد پارامترهای قابل یادگیری برای چند روش متداول و روش پیشنهادی را نشان می‌دهد. مجموع پارامترهای محاسبه‌شده برای هر لایه تعداد کل

[5] Stochastic Gradient Descent
[6] Momentum

پارامترهای قابل یادگیری شبکه را نمایش می‌دهد.

بهبود عملکرد این روش به‌ویژه در تقسیم‌بندی اشیا با تعداد پیکسل‌های کم، مانند فنس، ستون چراغ، دوچرخه‌سوار و تابلو علائم چشمگیر است. علت آن به دلیل عمق کم شبکه و استفاده کمتر از لایه max pooling است. زیرا این لایه در واقع نوعی کاهش نمونه در نقشه ویژگی‌هاست که باعث از بین رفتن اطلاعات و وضوح تصویر می‌شود [۲۹]. این کاهش تعداد پارامترها باعث کاهش حجم محاسبات و افزایش سرعت و کارایی شبکه برای کاربردهای برخط است به‌علاوه هزینه‌ها به علت کاهش نیاز به حافظه کاهش می‌یابد.

جدول ۱. مقایسه تعداد پارامترهای تنظیمی روش پیشنهادی و سایر روش‌ها در تقسیم‌بندی معنایی

نام معماری	نوع	تعداد پارامترها
SegNet [۲۰]	کانولوشن	۱۴/۷ مگا
ENet [۲۸]	کانولوشن	۰/۳۶ مگا
SqueezeNet [۲۹]	کانولوشن	۲/۷ مگا
VGG 16 [۳۰]	کانولوشن	۱۳۸ مگا
روش پیشنهادی	کانولوشن	۱/۴۱ مگا

سرعت همگرایی شبکه پیشنهادی نسبت به سایر روش‌ها در پایگاه داده CamVid بعد از ۴۰ هزار تکرار، در جدول (۲) نشان داده شده است [۲۲]. همان‌طور که مشاهده می‌شود روش پیشنهادی پس از تعداد تکرار کمتری به‌دقت مطلوبی نسبت به سایر روش‌ها دست یافته است. بنابراین سرعت شبکه پیشنهادی به‌طور قابل‌ملاحظه‌ای نسبت به سایر الگوریتم‌های موجود افزایش نشان می‌دهد.

جدول ۲. مقایسه عملکرد کلی روش پیشنهادی در تقسیم‌بندی معنایی با سایر روش‌ها برای پایگاه داده CamVid بعد از ۴۰ هزار تکرار

نام معماری	GA	CA	mIoU
SegNet [۲۰]	۸۸/۸۱	۵۹/۹۳	۵۰/۰۲
DeepLab-LargeFOV [۳۱]	۸۵/۹۵	۶۰/۴۱	۵۰/۱۸
FCN [۲۷]	۸۱/۹۷	۵۴/۳۸	۴۶/۵۹
FCN (learnt deconv) [۲۷]	۸۳/۲۱	۵۶/۰۵	۴۸/۶۸
DeconvNet [۳۲]	۸۵/۲۶	۴۶/۴۰	۳۹/۶۹
روش پیشنهادی	۸۹/۴۹	۸۳/۷۶	۶۲/۶۵

جدول (۳)، مقادیر معیارهای مقایسه‌ای جدول (۲) را با حداکثر تعداد تکرار برای بهترین پاسخ، مطابق با مقاله SegNet نشان می‌دهد. در این جدول مقایسه‌ای بین سایر روش‌ها و روش پیشنهادی از نظر بیش‌ترین تعداد تکرار برای دستیابی به بهترین پاسخ ارائه شده است. این در حالی است که روش پیشنهادی تنها برای ۶۰ هزار تکرار آموزش دیده است. بنابراین، الگوریتم پیشنهادی توانسته است با تعداد تکرار بسیار کمتری نسبت به سایر روش‌ها عملکرد مناسب‌تری را از خود نشان دهد که نشان‌دهنده سرعت همگرایی و دقت مناسب این روش است. نتایج نشان‌دهنده سرعت همگرایی و دستیابی به‌دقت بالاتر در تمامی معیارها نسبت به سایر روش‌هاست. مسلماً اگر تعداد تکرارهای آموزش افزایش یابد می‌توان حتی به نتایج بهتری دست یافت.

جهت بررسی شبکه پیشنهادی باید دقت تخمین جریان نوری با سایر روش‌های مقایسه شود. این مقایسه عملکرد در جدول‌های (۴) و (۵) برای پایگاه داده KITTI 2015 نشان داده شده است.

جدول ۳. مقایسه عملکرد کلی روش پیشنهادی و سایر روش‌ها برای پایگاه داده CamVid با افزایش تکرار آموزش

نام معماری	GA	CA	mIoU	تعداد تکرار ×۱۰۰۰
SegNet [۲۷]	۱۱/۸۱	۵۹/۹۳	۵۰/۰۲	۱۴۰
DeepLab-LargeFOV [۳۱]	۸۵/۹۵	۶۰/۴۱	۵۰/۱۸	۱۴۰
FCN [۲۲]	۸۱/۹۷	۵۴/۳۸	۴۶/۵۹	۲۰۰
FCN (learnt deconv) [۲۲]	۸۳/۲۱	۵۶/۰۵	۴۸/۶۸	۱۶۰
DeconvNet [۳۳]	۸۵/۲۶	۴۶/۴۰	۳۹/۶۹	۲۶۰
روش پیشنهادی	۹۱/۱۸	۸۴/۶۴	۶۵/۹۴	۶۰

جدول ۴. مقایسه عملکرد کلی روش پیشنهادی و سایر روش‌ها برای پایگاه داده KITTI 2015 و معیار AEE

نام روش	AEE
CPM-Flow [۳۴]	۷/۷۸
RIC Flow [۳۵]	۷/۲۴
CPM-OIR [۳۶]	۷/۳۶
DF-OIR [۳۶]	۵/۸۹
Pro-Flow [۳۷]	۵/۲۲
روش پیشنهادی	۵/۱۹

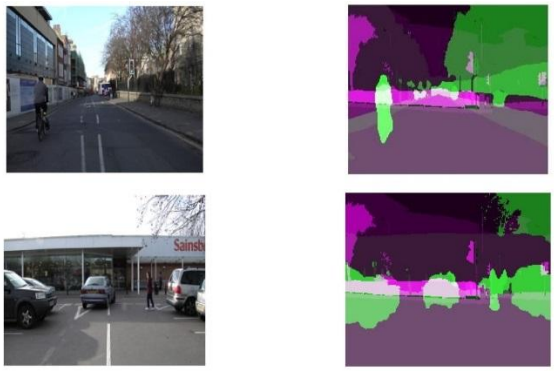
ازآنجاکه معیارهای عنوان‌شده نشان‌دهنده خطا در تخمین جریان نوری است بنابراین، هر چه مقادیر این معیارها کمتر باشد نشان‌دهنده عملکرد بهتر روش موردنظر است. نتایج جدول (۴) نشان می‌دهد که روش پیشنهادی کمترین میزان خطای نقطه انتهایی جریان (AEE) را در مقایسه با روش‌های دیگر دارد.

جدول ۵. مقایسه عملکرد کلی روش پیشنهادی و سایر روش‌ها برای پایگاه داده KITTI 2015 و معیار Fl-all

نام روش	Fl-all (%)
PWC-Net [۲۶]	۹/۶۰
Mirror Flow [۳۸]	۱۰/۲۹
Unflow [۳۹]	۱۱/۱۱
SOF [۱۳]	۱۶/۸۱
Pro-Flow [۳۷]	۱۵/۰۴
روش پیشنهادی	**۱۰/۵۸**

جدول (۵) نشان می‌دهد که روش پیشنهادی پاسخ مطلوبی نسبت به سایر روش‌ها به‌دست می‌دهد. روش پیشنهادی در مقایسه با روش‌های PWC-Net و Mirror Flow خطای بیشتری دارد در حالی که در مقایسه با این روش‌ها از ابعاد کوچک‌تر شبکه استفاده کرده است. به‌عنوان مثال شبکه PWC-Net دارای ۸,۴ مگا پارامتر قابل یادگیری است که بسیار بیشتر از روش پیشنهادی است [۲۶]. ابعاد کوچک شبکه باعث ایجاد قابلیت استفاده شبکه برای کاربردهای برخط می‌شود. بنابراین، با در نظر گرفتن ابعاد شبکه و معیار Fl-all روش پیشنهادی بازدهی بهتری را دارد. بنابراین، همان‌طور که در این جداول مشاهده می‌شود، روش پیشنهادی توانسته است به نتایج مطلوبی در زمینه تخمین جریان نوری دست یابد. در واقع استفاده از تقسیم‌بندی معنایی توانسته است به تخمین جریان نوری کمک کند. به‌علاوه چون ابعاد شبکه پیشنهادی کوچک است، سرعت همگرایی بالایی دارد.

در شکل (۹) نمایی از اجرای شبکه تقسیم‌بندی معنایی در روش پیشنهادی برای تصاویری از پایگاه داده CamVid نمایش داده شده است.



شکل ۹. خروجی شبکه تقسیم‌بندی معنایی روش پیشنهادی برای تصاویری از پایگاه داده CamVid

۴. نتیجه‌گیری

در این پژوهش یک رویکرد جدید تخمین جریان نوری با استفاده از شبکه‌های CNN رمزگذار-رمزگشا معرفی شد. شبکه پیشنهادی از دو بخش تقسیم‌بندی معنایی و تخمین جریان نوری تشکیل شده که به‌صورت متوالی قرار گرفته‌اند. در واقع هرکدام از دو فریم متوالی ابتدا به شبکه تقسیم‌بندی معنایی وارد شده و سپس نتیجه تقسیم‌بندی معنایی این دو فریم با یکدیگر جمع می‌شوند و به‌عنوان ورودی به شبکه تخمین جریان اعمال می‌شوند. با توجه به ابعاد کوچک شبکه پیشنهادی، این شبکه سرعت همگرایی مطلوبی از خود نشان می‌دهد و نیز نتایج در هر دو بخش تقسیم‌بندی معنایی و تخمین جریان نوری، نشان‌دهنده دقت مطلوب عملکرد این شبکه در هر دو مرحله تقسیم‌بندی معنایی و تخمین جریان نوری برای پایگاه داده‌های CamVid و KITTI 2015 هستند.

۵. مرجع‌ها

[1] Revaud, J.; Weinzaepfel, P.; Harchaoui, Z., Schmid, C. “Epic Flow: Edge-Preserving Interpolation of Correspondences for Optical Flow”; IEEE Conf. Comput. Vision Pattern Recgn. 2015, 1164-1172.

[2] Sun, D., Roth, S., Black, M. J. “A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them”; Int. J. Comput. Vision 2014, 106, 115-137.

[3] Butler, D. J.; Wulff, J.; Stanley, G. B.; Black, M. J. “A Naturalistic Open Source Movie for Optical Flow Evaluation”; European Conf. Computer Vision 2012, 7577, 611-625.

[4] Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. “Vision Meets Robotics: The KITTI Dataset”; Int. J. Robot. Res. 2013, 32, 1231-1237.

[5] Yamaguchi, K.; McAllester, D. A.; Urtasun, R. “Robust Monocular Epipolar Flow Estimation”; Proc. CVPR IEEE 2013, 1862–1869.

[6] Yamaguchi, K.; McAllester, D. A.; Urtasun, R. “Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation”; European Conf. Computer Vision 2014, 8693, 756-771.

[7] Lucas, B.; Kanade, T. “An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA)”; Proc. DARPA Image Understanding Workshop 1981, 121-130.

[8] Horn, B. K. P.; Schunk, B. G. “Determining Optical Flow”; Artif. Intell. Rev. 1981, 17, 185-203.

[9] Papenberg, N.; Bruhn, A.; Brox, T.; Didas, S.; Weickert, J. “Highly Accurate Optic Flow Computation with Theoretically Justified Warping”; Int. J. Comput. Vision 2006, 67, 141-158.

[10] Yang, H.; Lin, W.; Lu, J. “DAISY Filter Flow: A Generalized Discrete Approach to Dense Correspondences”; IEEE Conf. Comput. Vision Pattern Recgn. 2014.

[11] Bao, L.; Yang, Q.; Jin, H. “Fast Edge-Preserving Patch Match for Large Displacement Optical Flow”; IEEE Trans. Image Process. 2014, 23, 4996-5006.

[12] Menze, M.; Heipke, C.; Geiger, A. “Discrete Optimization for Optical Flow”; German Conf. Pattern Recogn. 2015, 9358, 16-28.

for Optical Flow Using Pyramid, Warping, and Cost Volume"; IEEE Conf. Comput. Vision Pattern Recgn. 2018

[27] Shelhamer, E.; Long, J.; Darrell, T. "Fully Convolutional Networks for Semantic Segmentation"; IEEE Trans. Pattern Anal. 2017, 39, 640–651.

[28] Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation"; arXiv preprint arXiv: 1606.02147, 2016.

[29] Nanfack, G.; Elhassouny, E.; Thami, R. O. H. "Squeeze-SegNet: A New Fast Deep Convolutional Neural Network for Semantic Segmentation"; Tenth Int. Conf. Machine Vision, 2017.

[30] Simonyan, K.; Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition"; arXiv Preprint arXiv: 1409.1556, 2014.

[31] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. "Deep Lab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs"; IEEE Trans. Pattern Anal. 2018, 40, 834-848.

[32] Noh, H.; Hong, S.; Han, B. "Learning Deconvolution Network for Semantic Segmentation"; IEEE Int. Conf. Comput. Vision 2015, 1520-1528.

[33] Tighe, J.; Lazebnik, S. "Super Parsing: Scalable Nonparametric Image Parsing with Super Pixels"; European Conference on Computer Vision 2010, 352-365.

[34] Hu, Y.; Song, R.; Li., Y. "Efficient Coarse-to-fine Patch Match for Large Displacement Optical Flow"; IEEE Conf. Comput. Vision Pattern Recgn. 2016, 5704-5712.

[35] Hu, Y.; Li, Y.; Song, R. "Robust Interpolation of Correspondences for Large Displacement Optical Flow"; IEEE Conf. Comput. Vision Pattern Recgn. 2017, 4791–4799.

[36] Maurer, D.; Stoll, M.; Bruhn, A. "Order-Adaptive and Illumination-Aware Variational Optical Flow Refinement"; Proc. of the British Machine Vision Conference 2017.

[37] Maurer, D.; Bruhn, A. "ProFlow: Learning to Predict Optical Flow"; arXiv preprint arXiv:1806.00800. 2018.

[38] Hur, J.; Roth, S. "Mirror Flow: Exploiting Symmetries in Joint Optical Flow and Occlusion Estimation"; IEEE Conf. Comput. Vision Pattern Recgn. 2017, 312-321.

[39] Meister, S.; Hur, J.; Roth, S. "Unflow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss"; Proc. AAAI Conf. Artificial Intelligence 2018.

[13] Yang, J.; Li, H. "Dense, Accurate Optical Flow Estimation With Piecewise Parametric Model"; IEEE Conf. Comput. Vision Pattern Recgn. 2015, 1019-1027.

[14] Sun, D.; Liu, C.; Pfister, H. "Local Layering for Joint Motion Estimation and Occlusion Detection"; IEEE Conf. Comput. Vision Pattern Recgn. 2014, 1098-1105.

[15] Sevilla-Lara, L.; Sun, D.; Jampani, V.; Black, M. J. "Optical Flow with Semantic Segmentation and Localized Layers"; IEEE Conf. Comput. Vision Pattern Recgn. 2016, 3889-3898.

[16] Farsi H.; Behmadi, S. "Video Quality Improvement Using Local Channel Encoder and Mixed Predictor by Wavelet, Neural Network and Genetic Algorithm"; J. Adv. Defense Sci. Technol. 2018, 9, 449-459.

[17] Zbontar, J.; LeCun, Y. "Computing the Stereo Matching Cost with a Convolutional Neural Network"; IEEE Conf. Comput. Vision Pattern Recgn. 2015, 1592–1599.

[18] Luo, W.; Schwing, A. G.; Urtasun, R. "Efficient Deep Learning for Stereo Matching"; IEEE Conf. Comput. Vision Pattern Recgn. 2016, 5695–5703.

[19] Geiger, A.; Lenz, P.; Urtasun, R. "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite"; IEEE Conf. Comput. Vision Pattern Recgn. 2012.

[20] Badrinarayanan, V.; Kendall, A.; Cipolla, R. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation"; IEEE Trans. Pattern Anal. 2017, 39, 2481-2495.

[21] Chantas, C.; Gkamas, T.; Nikou, C. "Variational-Bayes Optical Flow"; Journal of Mathematical and Imaging Vision 2014, 50, 199-213.

[22] Brostow, G. J.; Fauqueur, J.; Cipolla, R. "Semantic Object Classes in Video: A High-Definition Ground Truth Database"; Pattern Recogn. Lett. 2009, 30, 88-97.

[23] Tan, Z.; Liu, B.; Yu, N. "PPEDNet: Pyramid Pooling Encoder-Decoder Network for Real-Time Semantic Segmentation"; Int. Conf. Image and Graphics 2017, 328-339.

[24] Everingham, M.; Eslami, S. M. A.; Van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. "The Pascal Visual Object Classes Challenge: A Retrospective"; Int. J. Computer Vision 2015, 111, 98-136.

[25] Sharmin, N.; Brad, R. "Optimal Filter Estimation for Lucas-Kanade Optical Flow"; Sensors 2012, 12, 12694-12709.

[26] Sun, D.; Yang, X.; Liu, M. Y.; Kautz, Y. "PWC-Net: CNNs