بخش‌بندی معنایی بی‌ناظر تصاویر RGB-D با استفاده از ترکیب روش برش گراف و میدان تصادفی شرطی

سیدسعید میرکمالی*

استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور، تهران، ایران*

چکیده

هدف بخش‌بندی معنایی، اختصاص برچسب متناسب به مجموعه‌ای از پیکسل‌های یک شیء در یک تصویر با توجه به مشخصات ظاهری و معنایی آن است. این مسئله یکی از چالش برانگیزترین کارها در علم پردازش تصویر و بینایی ماشین و در سال‌های اخیر بسیار مورد توجه جامعهٔ بینایی ماشین قرار گرفته است. در این مقاله، روشی برای بخش‌بندی معنایی تصاویر RGB-D به‌صورت لایه‌به‌لایه ارائه شده‌است. الگوریتم پیشنهادی، ویژگی‌های ظاهری و اطلاعات عمق را در یک مدل میدان تصادفی شرطی (CRF) بدون نظارت یک‌پارچه می‌کند و از یک روش برش گراف کمک می‌گیرد تا یک صحنه را به لایه‌های منسجم و معنادار تقسیم کند. روش پیشنهادی از برش‌های گراف برای بهینه‌سازی فرایند برچسب‌گذاری استفاده می‌کند. در این مقاله برای ارزیابی عملکرد روش پیشنهادی از نظر کمی و کیفی از دو مجموعه‌دادهٔ مختلف استفاده شده‌است که هریک ویژگی‌های منحصربه‌فردی دارند؛ همچنین برای مقایسهٔ روش پیشنهادی نتایج به‌دست‌آمده با هشت روش بخش‌بندی معنایی نظارت‌شده و بدون ناظر دیگر مقایسه شده‌اند. نتایج آنالیزها نشان می‌دهد که CRF بدون نظارت می‌تواند به‌اندازهٔ روش‌های نظارت‌شده دقیق باشد و در بسیاری از موارد حتی می‌تواند بهتر از سایر روش‌های بخش‌بندی عمل کند.

واژگان کلیدی: بخش‌بندی لایه‌به‌لایه، تصویر RGB-D، میدان تصادفی شرطی CRF، برش گراف.

Unsupervised Semantic Segmentation of RGB-D Images Using Combination of Conditional Random Field with Graph Cuts

Seyedsaeid Mirkamali*

Assistant Professor Department of Computer Engineering and IT, Payame Noor University, Tehran, Iran*

Abstract

Semantic segmentation seeks to give a set of pixels depicting an object in an image suitable labels depending on their appearance and semantic characteristics. Though it is still one of the most difficult issues in image processing and computer vision, this work has attracted a lot of interest recently.

The availability of RGB-D sensors has introduced new possibilities for segmentation by incorporating depth information alongside color. However, effectively combining these modalities presents challenges due to misalignments and depth inaccuracies. This paper proposes CRFCut, a novel unsupervised segmentation method that utilizes a Conditional Random Field (CRF) model optimized with graph cuts to segment RGB-D images into coherent regions. The method recursively divides regions into foreground and background layers, employing superpixel-based appearance segmentation for the RGB component and integrating depth cues to refine results. This approach enables robust segmentation, even in the presence of noisy or incomplete depth information.

The CRFCut algorithm begins by separating the depth image into foreground and background regions using a median depth threshold. This initial step requires no preprocessing and provides the basis for further segmentation. Simultaneously, the RGB image is segmented into superpixels using an appearance-based approach, such as the mean-shift algorithm. These superpixels and the depth regions are combined within a CRF model, where labels are assigned by minimizing the energy function using the graph-cut α-expansion algorithm. The algorithm is applied recursively to subdivided regions, allowing finer segmentation in a parallelizable manner.

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۳/۹/۷ • تاریخ پذیرش: ۱۴۰۴/۴/۳۰ • تاریخ انتشار: ۱۴۰۴/۱۲/۲۹ • نوع مطالعه: پژوهشی

The proposed method was evaluated on two datasets: the NYUv2 dataset and the MIT dataset. The NYUv2 dataset, which includes 1449 RGB-D images with annotated object classes, demonstrated the superior performance of CRFCut compared to five state-of-the-art segmentation techniques in Table ۳. In the MIT dataset, which provides human-labeled sequences of indoor and outdoor scenes, CRFCut achieved comparable or better results, even with depth maps generated from 2D images using existing estimation methods (Table 4). The RandIndex metric was used to evaluate segmentation accuracy, and qualitative results, as shown in Figures 3, 4 and 5, highlight CRFCut's robustness, particularly with noisy or imprecise depth data.

In summary, CRFCut introduces an unsupervised CRF-based approach that integrates RGB and depth information for accurate scene segmentation. By leveraging graph-cut optimization and a recursive structure, the method achieves high-quality segmentation results with minimal preprocessing. Despite some limitations, such as challenges in distinguishing adjacent objects with similar features, CRFCut offers a promising framework for segmentation of RGB-D images. Future work will address these limitations by incorporating supervised techniques and improving depth data quality for enhanced performance.

Keywords: Semantic Segmentation, RGB-D Image, Combination of Conditional Random Field, Graph Cuts.

۱-مقدمه

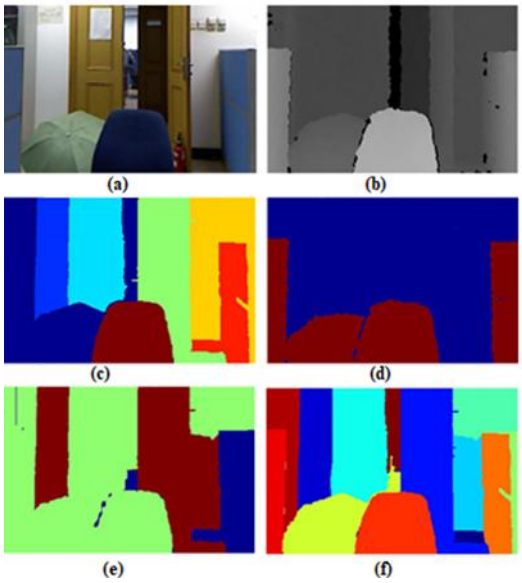
تقسیم خودکار یک تصویر به مناطق منسجم و معنی‌دار یکی از مهم‌ترین و دشوارترین مسائل بینایی ماشین است که کاربردهای فراوانی دارد؛ ازجمله کاربردهای بخش‌بندی معنایی تصاویر می‌توان به تشخیص و شمارش اجسام موجود در تصویر و یا حتی ویرایش خودکار تصاویر و اجسام موجود در آن اشاره کرد[۱۱، ۱۲]. بیشتر روش‌های بخش‌بندی بدون نظارت درکل به اطلاعات رنگ/بافت متکی‌اند [۱، ۲، ۴، ۱۳] که ممکن است برای دست‌یابی به یک نتیجهٔ بخش‌بندی خوب، کافی نباشد (شکل (c.۱) و شکل (f.۱) را مقایسه کنید). بسیاری از روش‌های بخش‌بندی [۱۵، ۱۶] از یک مجموعهٔ آموزشی برای یادگیری ویژگی‌های شی استفاده می‌کنند و در ادامه از آن‌ها برای کار بخش‌بندی استفاده می‌کنند. برخی روش‌های دیگر [۵، ۱۷-۱۹] از جریان نوری قاب‌های متعدد صحنه، برای حل این مسئله استفاده می‌کنند. استفاده از نماهای متعدد از یک صحنه راه دیگری برای تقسیم یک تصویر به مناطق منسجم و معنادار [۷، ۲۰] و ازجمله رویکردهای بسیار وقت‌گیر است. به‌تازگی، کم‌هزینه‌بودن و ماهیت زمان واقعی[1] دوربین‌های عمق، این امکان را برای پژوهش‌گران فراهم کرده‌است که روی تصاویر RGB-D[2] به‌عنوان ورودی برای الگوریتم‌های بخش‌بندی کار کنند[۲۱، ۲۲]، اما جفت‌کردن اطلاعات رنگ و عمق برای دست‌یابی به نتایج بخش‌بندی دقیق‌تر، کاری پرهزینه است. یک رویکرد ساده این است که فرضی ساده در مورد اتصال نقاط همسایه در تصویر عمق در نظر گرفته و تصویر RGB بر اساس تفاوت بین عمق آن‌ها برش داده شود، اما این روش به‌دلیل نبودِ تطابق مناسب بین RGB و تصویر عمق و در برخی موارد اطلاعات عمق ناکافی، منجر به بخش‌بندی نادرست می‌شود.

این مقاله با معرفی یک مدل CRF بدون نظارت با یک الگوریتم بهینه‌سازی بسط آلفا مبتنی بر برش گراف، به جفت‌کردن تصاویر RGB و عمق می‌پردازد[۹]. شکل (۱) نمای کلی از روش بخش‌بندی لایه‌ای پیشنهادی را نشان می‌دهد. روش پیشنهادی با تقسیم تصویر عمق به دو بخش پیش‌زمینه و پس‌زمینه، بدون هیچ‌گونه پیش‌پردازشی شروع می‌شود؛ سپس تصویر RGB را با استفاده از روش بخش‌بندی بدون نظارت به سوپرپیکسل‌ها تقسیم می‌کند. در ادامه از مدل CRF برای ایجاد برچسب‌های بالقوه از تصاویر RGB و عمق جهت بخش‌بندی استفاده می‌کند؛ علاوه‌براین، روش پیشنهادی از تابع انرژی متشکل از برچسب‌های داوطلب استفاده می‌کند تا برش‌های گراف را به کمترین حد برساند. این دنباله از عملیات به‌صورت بازگشتی بر روی هر منطقهٔ جداشده اعمال می‌شود تا زمانی که دیگر شرایط جداسازی وجود نداشته باشد. نوآوری‌های اصلی این مقاله به‌صورت خلاصه به‌شرح زیر است:

- تعریف مدل CRF بدون نظارت: استفاده از برچسب‌های باینری حاصل از تقسیم اولیهٔ تصویر عمق به پیش‌زمینه و پس‌زمینه به‌عنوان برچسب‌های ازپیش تعریف‌شده برای مدل CRF بدون نیاز به مجموعهٔ آموزشی؛
- بهینه‌سازی با برش گراف: یک‌پارچه‌سازی الگوریتم بسط آلفا مبتنی بر برش گراف برای کمینه‌سازی تابع انرژی CRF و دست‌یابی به مرزهای دقیق‌تر در تقسیم‌بندی؛
- ساختار بازگشتی لایه‌ای: طراحی یک الگوریتم بازگشتی لایه‌ای که امکان پردازش موازی زیرناحیه‌ها را فراهم کرده و مقیاس‌پذیری و کارایی بالاتری ارائه می‌دهد؛
- ترکیب تطبیقی اطلاعات RGB و عمق: ادغام هوشمند ویژگی‌های ظاهری سوپرپیکسل‌های RGB با داده‌های عمق جهت تحمل نوفه و نبود دقت در نقشه‌های عمق.

در ادامه ساختار مقاله به‌صورت زیر سازماندهی شده‌است: بخش دوم کارهای مرتبط انجام‌شده در پژوهش‌های پیشین مورد بررسی قرار گرفته و بلافاصله در بخش سوم الگوریتم پیشنهادی مورد بحث قرار گرفته است. در بخش چهارم مقایسهٔ نتایج اعمال الگوریتم پیشنهادی بر روی دو مجموعه‌داده در مقابل الگوریتم‌های به‌روز دیگر مورد بررسی قرار گرفته و درنهایت در بخش پنجم جمع‌بندی و پیشنهادهایی جهت پژوهش‌های آینده ارائه شده‌است.

[1] Real Time
[2] A RGB image with Depth information



(شکل- ۱): نمای کلی از روش بخش¬بندی لایه‌ای پیشنهادی را نشان می‌دهد. (a) تصویر اصلی (RGB) دریافت‌شده از دوربین. (b) تصویر عمق. (c) تصویر بخش‌بندی‌شده بر اساس ویژگی‌های ظاهری. (d) تصویر بخش‌بندی‌شده بر اساس داده‌های عمق. (e) ترکیب ویژگی‌های ظاهری و عمق جهت انجام بخش‌بندی تصویر. (f) هر رنگ یک شی را در لایهٔ مربوطهٔ خود به‌عنوان نتیجهٔ نهایی روش پیشنهادی پس از بهینه‌سازی نشان می‌دهد.

(Figure -1): An overview of the proposed layered segmentation method is presented. (a) The original (RGB) image captured by the camera. (b) The depth image. (c) The segmented image based on appearance features. (d) The segmented image based on depth data. (e) The combination of appearance features and depth for image segmentation. (f) Each color represents an object in its respective layer as the final result of the proposed method after optimization.

۲- کارهای مرتبط

بیشتر رویکردهای بخش‌بندی لایه‌ای یک لایه را برای هر جسم متحرک فرض می‌کنند که تمام لایه‌های یک تصویر به‌ویژه اشیای ساکن را توصیف نمی‌کند؛ آن‌ها همچنین روی جداسازی یک صحنه به لایه‌های هم‌پوشانی، استدلال در مورد تعداد لایه‌ها و تعیین ترتیب عمق کار می‌کنند[۵، ۲۳]. تعدادی از روش‌ها روی نشانه‌های تصویر ایستا برای غلبه بر مسئلهٔ بخش‌بندی لایه‌ای کار می‌کنند[۲۴].

سیلبرمن و همکاران [۲۱] به‌منظور تقسیم یک تصویر با استفاده از تصویر عمق متناظر آن، روشی را پیشنهاد کرد که با الگوریتم حوضه[1] شروع می‌شود تا تصویر RGB را تا جای ممکن تقسیم کند؛ سپس الگوریتم مکرر جفت مناطق مجاور را بر اساس شباهت‌های آموخته‌شده ادغام می‌کند. برای طبقه‌بندی دو ناحیه برای نمونه، شیء یکسان براساس نشانه‌هایی از تصویر RGB و تصویر عمق، در این الگوریتم یک درخت تصمیم تقویت‌شدهٔ آموزش‌دیده طراحی شده‌است. نقطهٔ ضعف اصلی این الگوریتم این است که روش پیشنهادی یک روش نظارت‌شدهٔ وابسته به تعداد کمابیش زیادی نمونه برای هر شی در مجموعهٔ آموزشی است؛ در مقابل، روش ما یک روش بدون نظارت است که تنها از اطلاعات RGB و عمق استفاده می‌کند.

در سال‌های اخیر روش Potts MRF [۲۵، ۲۶] برای حل مشکل بخش‌بندی لایه‌ای به‌دلیل دردسترس‌بودن بهینه‌سازهای کارآمد مانند برش‌های گراف [۱۵] مورد توجه قرار گرفته‌است. میدان‌های تصادفی شرطی یکی از مدل‌های محبوب Potts MRF است. به‌تازگی انواع مدل‌های CRF برای حل مسائل برچسب‌گذاری ارائه شده‌اند[۶، ۲۷]. مدل CRF نزدیک به مدل ارائه‌شده را در این مقاله کوهلی و همکاران [۲۷] ارائه کرده‌اند. آن‌ها مشکل برچسب‌گذاری را همان‌طور که یک مدل CRF بر اساس سوپرپیکسل‌ها[2] (قطعات تصویر) تعریف می‌کند، فرموله کرده‌اند. آن‌ها از طیفی از سوپرپیکسل‌ها با اندازه‌های مختلف برای اعمال ثبات برچسب در مناطق تصویر در مجموعهٔ آموزشی استفاده می‌کنند. در مقابل روش پیشنهادی از سوپرپیکسل‌های تصویر RGB و احتمال گروه‌بندی آن‌ها به‌عنوان مناطق اولیه استفاده می‌کند. کوهلی و همکاران [۲۷] نشان دادند که توابع انرژی متشکل از برچسب‌های پتانسیل را می‌توان با استفاده از برش گراف بر اساس الگوریتم بسط آلفا به کمینه رساند. روش پیشنهادی همچنین از همان روش کمینه‌سازی برای بهینه‌سازی نتایج بخش‌بندی لایه‌ای استفاده کرده‌است.

دوو و همکارانش [۱۴] جهت درک تصاویر محیط‌های داخلی برای مطالعات شهری با توجه به ماهیت پویای محیط‌های داخلی، روش تقسیم‌بندی معنایی AsymFormer را پیشنهاد کردند. آن‌ها شبکهٔ جدیدی ارائه دادند که دقت تقسیم‌بندی معنایی بلادرنگ را با استفاده از اطلاعات چندوجهی RGB-D بدون افزایش قابل‌توجهی پیچیدگی شبکه، بهبود می‌بخشد. روش پیشنهادی از یک ستون نامتقارن برای استخراج ویژگی‌های چندوجهی استفاده می‌کند و با بهینه‌سازی توزیع منابع محاسباتی، مؤلفه‌های اضافی را کاهش می‌دهد.

ایین و همکارانش [۲۸] با استفاده از DFormer، یک چهارچوب پیش‌آموزشی جدید RGB-D را برای یادگیری نمایش‌های قابل انتقال برای وظایف تقسیم‌بندی RGB-D ارائه کردند. DFormer دو نوآوری کلیدی جدید دارد: ۱) برخلاف کارهای پیشین که اطلاعات RGB-D با اتکا به RGB ازپیش آموزش‌داده‌شده رمزگذاری می‌شدند، DFormer بیشتر متکی است به استفاده از جفت‌های عمق تصویر که ازپیش آموزش داده شده‌اند؛ ازاین‌رو DFormer دارای ظرفیت رمزگذاری RGB-D است. ۲) DFormer شامل دنباله‌ای از بلوک‌های RGB-D است که برای رمزگذاری اطلاعات RGB و عمق از طریق یک طراحی جدید بلوک‌های ساختمانی طراحی

[1] Watershed Algorithm

2 Superpixels

شده‌اند. DFormer از کدگذاری ناهمگن روابط هندسهٔ سه‌بعدی در نقشه‌های عمقی به‌وسیلهٔ نمونه‌های ازپیش آموزش‌دیده RGB جلوگیری می‌کند، که به‌طور گسترده در روش‌های پیشین استفاده شده‌است، اما مشکل به‌صورت جدی حل نشده‌است. آن‌ها DFormer ازپیش آموزش‌دیده را در دو وظیفهٔ اصلی RGB-D، یعنی تقسیم‌بندی معنایی RGB-D و تشخیص اشیای برجسته RGB-D، با یک روش رمزگشای آسان، تنظیم کردند.

در [۲۹] یک شبکهٔ هم‌جوشی مبتنی بر توجه برای تقسیم‌بندی معنایی RGB-D پیشنهاد شده‌است. شبکهٔ پیشنهادی از یک استراتژی انتشار چندمرحله‌ای روبه‌جلو و یک استراتژی ترکیبی روبه‌عقب بر اساس معماری رمزگذار-رمزگشا استفاده می‌کند. نویسندگان با جمع‌آوری نقشه‌های ویژگی در مقیاس‌های مختلف، به‌طور مؤثر عدم قطعیت در پیش‌بینی نهایی را کاهش دادند. آن‌ها یک کانال و ماژول اصلاح فضایی (CSRM) را برای فعال‌کردن تعاملات چندبعدی و حذف نوفه معرفی کردند؛ همچنین به‌منظور دست‌یابی به یک‌پارچگی جامع بین تصاویر RGB و عمق، ویژگی‌های اصلاح‌شده را در ماژول فیوژن متقاطع (CAFM) قرار دادند.

در [۳۰] یک شبکهٔ جدید انتخاب ویژگی هدایت‌شده با اولویت عمق[1] پیشنهاد شده‌است که اطلاعات عمق را به‌عنوان اطلاعات اولیه دریافت می‌کند و به‌صورت پویا اطلاعات RGB مکمل را برای تشخیص شی برجسته RGB-D انتخاب می‌کند؛ به‌طور خاص، DGFSnet ابتدا شامل یک ماژول تولید وزن عمق به نام DWG می‌شود تا مجموعه‌ای از وزن‌های خاص لایه را از ویژگی‌های عمق چندمقیاسی یاد بگیرد. با هدایت این وزن‌های آموخته‌شده، DGFSnet یک ماژول جمع‌آوری ویژگی هدایت‌شده با وزن (WFA) را ارائه می‌کند تا آن‌ها را به لایه‌های RGB متناظر خود برای افزایش پویا و انتخاب ویژگی‌های RGB مرتبط با برجسته‌بودن اختصاص دهد. روش ارائه‌شده با دو ماژول DGFSnet قادر است به طور مؤثری مکمل‌های چندوجهی را ادغام کند و مناطق برجسته را بیشتر برجسته کند.

تانگ و همکارانش [۳۱] یک طرح هم‌جوشی چندسطحی هدایت‌شده با استفاده از هندسه را برای تخمین نرمال سطح با کیفیت بالا با بهره‌برداری از اطلاعات بافت و هندسه از تصاویر رنگی و عمقی پیشنهاد کردند. سطح نرمال به‌تدریج با یک استراتژی درشت به ریز پیش‌بینی می‌شود. در این کار، آن‌ها بر برآورد نرمال سطح از داده‌های RGB-D تمرکز و یک طرح ترکیبی RGB-D چندسطحی را پیشنهاد کردند. ماژول FF اطلاعات بافت و هندسهٔ صحنه را در سطح ویژگی CNN ادغام و همچنین یک IniNormal با جزئیات واضح و صاف در سطح عمده ایجاد می‌کند. ماژول SNF بر نواحی خرابی در IniNormal غلبه خواهد کرد و اطلاعات را در سطح نرمال سطحی دوباره ادغام می‌کند؛ درکل، روش آن‌ها بارها اطلاعات هندسی تصویر را در نمایش‌های رمزگذاری مختلف نقشه‌برداری خواهد کرد.

ژیونگ و همکاران [۳۲] یک معماری یادگیری عمیق به‌نام AGWNet را معرفی کردند که برای تجزیهٔ تصاویر مناطق داخل ساختمان با استفاده از داده‌های RGB-D طراحی شده‌است. رمزگذار AGWNet دارای یک ماژول جدید به نام AGFM است که سازوکارهای دروازه‌ای تعاملی بین ویژگی‌های دوگانه را افزایش می‌دهد؛ علاوه‌بر این، رمزگذار ویژگی‌ها را به یک هرم تصحیح چند مقیاسی هدایت می‌کند که از هم‌سویی ویژگی‌های مودال برای انجام کار استفاده می‌کند. این استراتژی تصحیح ویژگی، به‌طور خاص در کاهش موارد غیرقابل پیش‌بینی بسیار خوب عمل می‌کند. جدول (۱) خلاصه و دسته‌بندی روش‌های توصیف‌شده را بیان می‌کند.

۳- الگوریتم بخش‌بندی

یک مدل CRF به مجموعه‌ای از برچسب‌های ازپیش تعریف‌شده برای ترکیب برچسب‌های بالقوه نیاز دارد؛ به‌طور معمول مجموعهٔ ازپیش تعریف‌شدهٔ برچسب‌ها با استفاده از یک مجموعهٔ آموزشی به‌دست می‌آید. در این مقاله، یک مدل CRF بدون نظارت تعریف شده است که برای یافتن برچسب‌های ازپیش تعریف‌شده نیازی به مجموعهٔ آموزشی ندارد. روش پیشنهادی از برچسب‌های تصویر عمق بخش‌بندی‌شده به‌عنوان یک برچسب باینری ازپیش تعریف‌شده استفاده می‌کند. در این بخش مدل CRF پیشنهادی و الگوریتم بخش‌بندی، به‌تفصیل توضیح داده خواهد شد.

۳-۱- مدل میدان تصادفی شرطی[2]

CRF یکی از روش‌های محبوب برای بهبود نتایج مدل‌های بخش‌بندی اشیا در تصاویر است. CRF یک مدل گراف احتمالی است که با استفاده از روابط محلی بین پیکسل‌ها، خروجی‌های شبکه‌های عصبی (مانند FCN یا U-Net) را بهبود می‌بخشد و با توجه به محتوای محلی (همسایگی پیکسل‌ها)، مرزهای دقیق‌تر و نتایج روان‌تری را ایجاد می‌کند. هدف استفاده از CRF اطمینان‌یافتن از این است که خروجی مدل نه‌تنها به‌درستی اشیا را پیش‌بینی می‌کند، بلکه مرزهای دقیق‌تری بین اشیای مختلف ایجاد می‌کند.

الگوی CRF پیشنهادی در این مقاله بر اساس گراف بدون جهت $G=(V,E)$ تعریف شده‌است که در آن گره‌ها V مجموعهٔ سوپرپیکسل‌های تصویر و یال‌ها E بیان‌کنندهٔ روابط همسایگی بین آن‌ها هستند. هدف مدل CRF، یافتن برچسب بهینهٔ $x=\{x_i\}$ برای هر گره $i\in V$ است؛ به‌طوری‌که

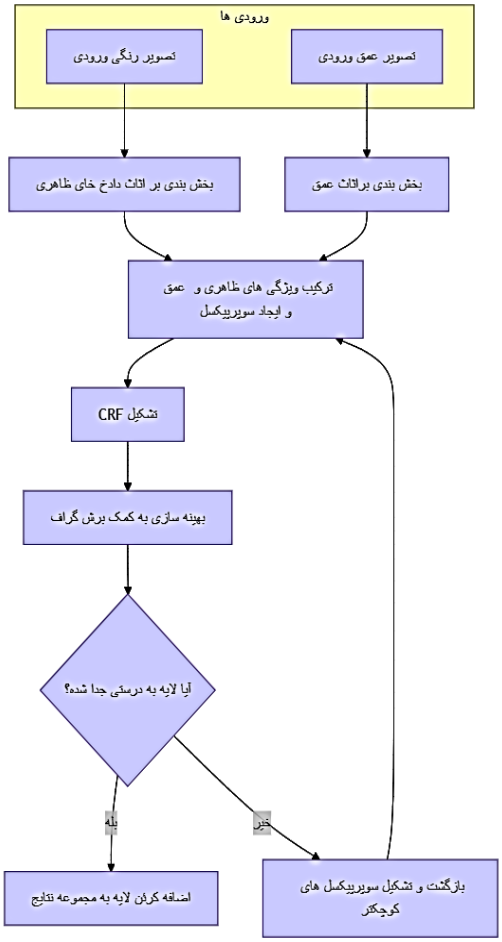
[1] DGFSnet

[2] Conditional Random Field (CRF)

تابع انرژی $E(y,x;\omega)$ کمینه شود. همان‌طور که مشخص است، برای استفاده از این روش ابتدا باید تصاویر بر اساس ویژگی‌های رنگ، موقعیت و عمق به بخش‌های کلی به‌نام سوپرپیکسل تبدیل شوند.

(شکل- ۲): نمودار جعبه‌ای روش پیشنهادی
(Figure- 2): Block diagram of the proposed method

در اینجا، همان‌طور که در الگوریتم (۱) خط (۱) بیان شد، هر سوپرپیکسل با استفاده از تغییر میانگین و تقسیم تصویر عمقی D به دو قسمت مقداردهی اولیه می‌شود و



به‌نام $x^{(n)}$. برچسب‌گذاری می‌شود؛ به‌دلیل ماهیت دودویی مجموعهٔ X، مدل پیشنهادی یک میدان تصادفی شرطی به‌شکل رابطهٔ (۱) خواهد بود[۱].

در اینجا ω وزن است (وزن‌های اختصاص‌داده‌شده به سوپرپیکسل‌ها بر اساس همگنی رنگ و بافت با استفاده از بخش‌بندی میانگین شیفت [۱]، Z یک نرمال‌ساز (تابع تقسیم x) است.

$$P(Y|X,\omega) = \frac{1}{\mathrm{z}(x)}\left(\frac{exp\{\sum_{x_i\in X}\omega_i x_i\}}{(1+exp\{\sum_{x_i\in X}\omega_i x_i\})}\right) \quad (۱)$$

در ادامه مانند سزومر و همکاران [۸] در مدل پیشنهادی از برچسب‌گذاری برآوردگر بیشینه‌گر احتمال پسین (MAP) که به‌صورت $y^* = argmax(y|x,\omega)$ تعریف‌شده‌است برای استنتاج مجموعه‌برچسب‌های یک مجموعهٔ ورودی X استفاده می‌کنیم. در ادامه MAP به یک مسئله برش نمودار به شکل رابطهٔ (۲) خواهد آمد.

$$E(y,x;\omega) = \sum_{i\in v} P(y_i|x_i,\omega) + \sum_{(i,j)\in\varepsilon}\Phi(x_i,x_j|y_i,y_j) \quad (۲)$$

در این رابطه، انرژی E را با مجموعه‌ای از گره‌ها v (پیکسل‌های منفرد) و ویژگی‌های لبه (جفت پیکسل‌های مجاور) در نمودار $\emptyset$ درنظر می‌گیریم که در آن $P(x_i)$ احتمال تک‌متغیره[1] است که هزینهٔ انتساب برچسب x_i به گره i را بر اساس شباهت با ویژگی‌های اولیه (مانند شدت رنگ، موقعیت و عمق) تعریف می‌کند. $\Phi(x_i,x_j)$ احتمال جفت‌متغیره[2] است که تنبیه تفاوت بین دو برچسب همسایه i و j را مدل می‌کند تا انسجام فضایی حفظ شود. در مدل پیشنهادی، تابع هزینهٔ تک‌متغیره با استفاده از برچسب اولیهٔ عمق دودویی (پیش‌زمینه/پس‌زمینه) به‌صورت تابع فاصله تعریف شده‌است و احتمال تعلق به پیش‌زمینه/پس‌زمینه بر اساس عمق و احتمال دومتغییری بر اساس ویژگی‌های رنگی در فضای سوپرپیکسلی تعریف شده‌است.

برای کمینه‌سازی تابع انرژی بالا، از الگوریتم بسط آلفا[3] استفاده شد که روشی کمابیش بهینه برای مسائل تخصیص چندبرچسبی است[۳۳]. در این الگوریتم، در هر گام برچسب آلفا به مجموعه‌ای از گره‌ها گسترش داده می‌شود و با استفاده از الگوریتم Max-Flow/Min-Cut روی گراف، تصمیم گرفته می‌شود که کدام گره‌ها تغییر برچسب دهند. براساس برش گراف پیشنهادشده توسط *بویکو و همکاران*[۹]، فرایند بسط آلفا تا زمانی ادامه می‌یابد که تغییر در انرژی کمتر از یک آستانهٔ ε باشد یا تغییرات ساختاری ناچیز شود.

۲-۳-الگوریتم پیشنهادی با استفاده از میدان تصادفی شرطی

همان‌طور که پیش‌تر بیان شد، الگوریتم پیشنهادی بر اساس یک مدل CRF بدون نظارت طراحی و پیاده‌سازی شده‌است. این الگوریتم تصویر رنگی (RGB) و همچنین تصویر عمق متناظر آن (D) را به‌عنوان ورودی دریافت می‌کند و تصویر را به لایه‌های شی منسجم و معنادار تقسیم می‌کند. شکل (۲) نمودار جعبه‌ای این مراحل را توصیف می‌کند و الگوریتم (۱) شبه‌کد الگوریتم بخش‌بندی لایه‌ای بازگشتی پیشنهادی در این مقاله را نشان می‌دهد. جدول (۲) مؤلفه‌های مورد استفاده در این روش را توصیف می‌کند.

[1] unary potential
[2] pairwise potential
[3] α − expansion

(جدول-۱): خلاصهٔ روش‌های مختلف بخش‌بندی تصاویر RGB-D که در کارهای مرتبط پیشین مورد بررسی قرار گرفته است.

(Table- 1) :Summary of various RGB-D image segmentation methods that have been reviewed in this paper.

دسته	روش	توضیحات کلیدی	مرجع
سوپرپیکسل و خوشه‌بندی	میانگین-شیفت [۱]	خوشه‌بندی بر اساس تغییر میانگین در فضای ویژگی	Comaniciu & Meer (2002)
	بخش‌بندی تصاویر بر پایهٔ گراف [۴]	برش گراف مبتنی بر وزن لبه‌ها برای تقسیم تصویر	Felzenszwalb & Huttenlocher (2004)
میدان تصادفی شرطی (CRF)	Associative Hierarchical CRFs [6]	پتانسیل‌های سلسله‌مراتبی برای حفظ انسجام برچسب	Ladický et al. (2009)
	CRF با برش گراف [۸]	یک‌پارچه‌سازی CRF با برش گراف برای بهینه‌سازی انرژی	Szummer et al. (2008)
برش گراف و بهینه‌سازی انرژی	بسط آلفا[۹]	الگوریتم بسط آلفا برای کمینه‌سازی تابع انرژی دوبخشی	Boykov et al. (2001)
	GrabCut [10]	استخراج پیش‌زمینه با استفاده از برش گراف و مدل GMM	Rother et al. (2004)
ترکیبی RGB-D و فیوژن چندوجهی	AsymFormer [14]	ساختار نامتقارن برای فیوژن چندوجهی دینامیک	Du et al. (2024)

(الگوریتم- ۱): شبه‌کد الگوریتم بخش‌بندی لایه‌ای بازگشتی

ورودی‌ها
- تصویر رنگی: تصویر $RGB_{n\times m}$
- عمق $D_{n\times m}$
- ناحیهٔ مورد علاقه: $\mathrm{I} = \{n \times m\}$
- مناطق برچسب‌گذاری شدهٔ نهایی: $\mathrm{S} = \emptyset$

الگوریتم
۱. $(D \cap I) \rightarrow x$ تقسیم به دو گروه با استفاده از (۳)
۲. $\mathrm{Y}\ (RGB \cap I) \rightarrow y$ بخش را به سوپرپیکسل‌ها تقسیم کن.
۳. تشکیل شبکهٔ CRF به کمک معادلهٔ (۱)
٤. برچسب‌گذاری MAP با استفاده از برش گراف (۲)
٥. $\mathrm{S} \cup \{y^*\} \rightarrow S$
٦. اگر $(D \cap I)$ تغییر کند (به‌اندازهٔ ۱۰درصد)
۷. حلقه روی تمام n ناحیهٔ S
   a) $S^{(n)} \rightarrow I$
   b) نواحی مجاور $(RGB \cap I)$و $(D \cap I)$ را هموارسازی کن.
   c) . الگوریتم را با مقادیر جدید تکرار کن، در I، RGB و D

الگوریتم با تصویر ورودی به‌عنوان ناحیهٔ اولیهٔ پیشنهادی برای بخش‌بندی شروع می‌شود. در اینجا $S = \emptyset$ نشان می‌دهد، مجموعهٔ نهایی مناطق تقسیم‌شدهٔ S در ابتدا خالی است. الگوریتم با تقسیم تصویر عمق به دو قسمت به‌عنوان برچسب‌های اولیه (X)؛ یعنی پیش‌زمینه و پس‌زمینه شروع می‌شود. دو نقطه با کمترین مقدار عمق و دیگری با بیشترین مقدار عمق برای شروع درنظر گرفته شد و جداسازی با استفاده از رابطهٔ (۳) انجام می‌شود.

$$X = \begin{cases} 0, & D_{i,j} < P \\ 1, & D_{i,j} \geq P \end{cases} \qquad (۳)$$

که در آن P یک نقطهٔ میانی بین بیشترین و کمترین مقدار تصویر عمق است، $D_{i,j}$ مقدار عمق پیکسل فعلی است. برای بخش‌بندی ناحیهٔ مورد نظر تصویر RGB به سوپرپیکسل، از روش بخش‌بندی مبتنی بر ظاهر استفاده می‌شود. انتخاب پژوهش‌گران مقالهٔ حاضر استفاده از روش بخش‌بندی میانگین شیفت [۱] با توجه به برتری میانگین تغییر در مطالعهٔ ارزیابی ارائه‌شده به‌وسیلهٔ شیء و مالیک [۱۳] است. با داشتن یک متغیر دودیی X و مجموعه‌ای از سوپرپیکسل‌های Y ، یک CRF به‌شکل رابطهٔ شمارهٔ (۱) تعریف می‌شود. انرژی رابطهٔ شمارهٔ (۱) ایجادشده در بخش ۳-۲ به‌کمک برش گراف براساس الگوریتم گسترش آلفا به کمینه رسانده می‌شود. کمینه‌سازی انرژی تضمین می‌کند که سوپرپیکسل‌های تصویر RGB براساس مقادیر عمق متناظرشان به‌طور دقیق به دو دسته تقسیم می‌شوند.

اگر میزان تغییر بین S و I قابل توجه باشد (بیش از ده درصد) روی هر دو زیر منطقهٔ S به‌صورت بازگشتی الگوریتم دوباره اعمال می‌شود. برای شروع با یک مسئلهٔ بخش‌بندی جدید، ناحیهٔ ورودی (I) براساس یک زیرمنطقهٔ ورودی S تغییر می‌کند.

(جدول-۲): مؤلفه‌های روش پیشنهادی.

(Table- 2): parameters of the proposed method.

پارامتر	توضیحات
ω	وزن‌های اختصاص‌داده‌شده به سوپرپیکسل‌ها بر اساس همگنی رنگ و بافت با استفاده از بخش‌بندی میانگین شیفت.
E	تابع انرژی: ترکیبی از احتمال شرطی و پتانسیل زوجی برای بهینه‌سازی برچسب‌ها.
P	آستانهٔ عمق: نقطهٔ میانی برای تقسیم تصویر عمق به پیش‌زمینه و پس‌زمینه.
σ	پارامتر هموارسازی برای ادغام نواحی مجاور.
α	تعداد تکرارهای الگوریتم بسط آلفا برای بهینه‌سازی انرژی.

۴- نتایج شبیه‌سازی

برای نشان‌دادن توانمندی و کیفیت عملکرد روش بخش‌بندی لایه‌ای پیشنهادی که CRFCut نام‌گذاری شده، در این مقاله دو مجموعه‌دادهٔ متفاوت درنظر گرفته شد. نخستین مجموعه‌داده‌ای که برای ارزیابی رویکرد پژوهش استفاده شد، مجموعه‌دادهٔ RGB-D NYUv2 [۲۱] است. مجموعه‌دادهٔ دوم، مجموعه‌دادهٔ برچسب‌گذاری‌شدهٔ MIT [۳] است که توالی‌هایی از تصاویر RGB را با حقیقت عینی[1] یا برچسب‌گذاری‌شده برای بخش‌بندی لایه‌ای هر فریم ارائه می‌دهد. برای مقایسهٔ کیفیت عملکرد الگوریتم‌هایی که از نوع بخش‌بندی معنایی تصاویرند به‌طورمعمول از سه پارامتر مهم استفاده می‌شود که در این مقاله نیز از آن‌ها استفاده شده‌است. برای آشنایی با چگونگی این سه پارامتر ابتدا روش محاسبهٔ هریک مورد بحث و بررسی قرار می‌گیرد و سپس در جدول (۳) مقایسهٔ روش پیشنهادی با پنج روش برتری که در سال‌های گذشته بر روی مجموعه‌دادهٔ NYYv2 آزمایش شده‌اند فهرست خواهد شد؛ همچنین شکل (۳) نمونه‌ای از نتایج به‌دست‌آمده از بخش‌بندی معنایی روش پیشنهادی را نشان می‌دهد.

دقت پیکسل[2] (PA): یکی از ساده‌ترین معیارها برای ارزیابی عملکرد مدل‌های تقسیم‌بندی اشیا است. این معیار نشان می‌دهد که چه درصدی از پیکسل‌ها به‌درستی به‌وسیلهٔ مدل دسته‌بندی شده‌اند. این معیار به‌طور کلی برای وظایف بخش‌بندی معنایی استفاده می‌شود و نسبت تعداد پیکسل‌های صحیح پیش‌بینی‌شده به کل پیکسل‌های تصویر را محاسبه می‌کند. PA تنها به تعداد کلی پیکسل‌هایی که به‌درستی دسته‌بندی شده‌اند توجه می‌کند و تفاوتی بین طبقه‌ها قائل نمی‌شود.

مراحل محاسبهٔ PA:

پیکسل‌های صحیح پیش‌بینی‌شده: تعداد کل پیکسل‌هایی را که مدل به‌درستی پیش‌بینی کرده‌است، محاسبه می‌کند. این پیکسل‌ها شامل پیکسل‌های تمام طبقات می‌شوند $(\sum_{j=1}^{k} njj)$.

کل پیکسل‌ها: تعداد کل پیکسل‌های موجود در تصویر (یا مجموعه‌داده) است. این تعداد شامل پیکسل‌های درست پیش‌بینی‌شده و نادرست پیش‌بینی‌شده‌است $(\sum_{j=1}^{k} t_j')$.

$$PA = \frac{\sum_{j=1}^{k} njj}{\sum_{i=1}^{k} t_i'} \qquad (۴)$$

میانگین دقت پیکسل[3] (mPA): از آنجایی که چندین طبقه در تقسیم‌بندی معنایی وجود دارد، میانگین دقت پیکسل (mPA)، متوسط طبقه به‌صورت رابطهٔ شمارهٔ (۵) محاسبه می‌شود:

$$mPA = \frac{1}{k}\sum_{j=1}^{k} \frac{njj}{t_j} \qquad (۵)$$

تقاطع وزنی فرکانس بر روی اجتماع[4] (FWIoU): معیاری است برای ارزیابی عملکرد مدل‌های تشخیص و تقسیم‌بندی اشیا، به‌ویژه در زمینهٔ وظایف تقسیم‌بندی معنایی. این معیار، نسخه‌ای توسعه‌یافته از معیار پایه‌ای شاخص ژاکارد[5] (IoU) است، اما به طبقه‌هایی که بیشتر در داده‌ها ظاهر می‌شوند، وزن بیشتری می‌دهد. FWIoU به‌ویژه زمانی مفید است که یک مجموعه‌دادهٔ نامتوازن وجود داشته باشد که در آن برخی از طبقات شایع‌تر از سایرین‌اند. با وزن‌دادن به IoU بر اساس فراوانی هر طبقه، FWIoU به درنظرگرفتن هم‌طبقه‌های رایج و هم‌طبقه‌های نادر در ارزیابی نهایی کمک می‌کند.

مراحل محاسبهٔ *FWIoU*

(IoU): ابتدا یادآوری می‌شود که IoU نسبت هم‌پوشانی بین بخش‌بندی پیش‌بینی‌شده و تقسیم‌بندی واقعی به اجتماع آن‌هاست. IOU برای هر طبقه c به‌صورت زیر تعریف می‌شود:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \qquad (۶)$$

در اینجا:

- **مثبت واقعی[6] (TP):** پیکسل‌هایی که به‌درستی به‌عنوان متعلق به طبقه c پیش‌بینی شده‌اند.

[1] Ground Truth
[2] Pixel Accuracy
[3] Mean Pixel Accuracy
[4] Frequency Weighted Intersection over Union
[5] Intersection over Union
[6] True Positive

(a) (b) (c) (d)

(شکل- ۳): مناطق ارزیابی برای برخی از تصاویر از مجموعه‌دادهٔ NYUv2، در هر ستون (a) تصاویر اصلی، (b) بخش‌بندی حاصل اجرای الگوریتم[۲] TransD-Fusion که بهترین عملکرد را بین مقالات بررسی‌شده داشته است، (c) نتایج اجرای روش پیشنهادی CRFCut، (d) تصاویر برچسب‌گذاری‌شده توسط انسان.

(Figure-3): Evaluation regions for some images from the NYUv2 dataset. In each column: (a) Original images, (b) Segmentation results from the algorithm [33] TransD-Fusion, which demonstrated the best performance among the reviewed studies, (c) Results from the proposed method CRFCut, (d) Images labeled by humans.

- **مثبت کاذب[1] (FP):** پیکسل‌هایی که به‌اشتباه به‌عنوان متعلق به طبقه c پیش‌بینی شده‌اند.
- **منفی کاذب[2](FN):** پیکسل‌هایی که متعلق به طبقه c هستند، اما به‌عنوان این طبقه پیش‌بینی نشده‌اند.

محاسبهٔ فراوانی طبقه: برای هر طبقه c، فراوانی آن طبقه در داده‌های واقعی محاسبه می‌شود:

$$F_c = \frac{\text{تعداد پیکسل‌های طبقه}}{\text{تعداد کل همهٔ پیکسل‌ها در همهٔ طبقه‌ها}} \quad (۷)$$

محاسبهٔ FWIoU: برای هر طبقه IoU در فراوانی آن طبقه در داده‌های واقعی ضرب؛ سپس این IoU های وزنی برای همهٔ طبقه‌ها جمع می‌شوند:

$$FWIoU = \sum_{c=1}^{C}(F_c \times \frac{TP_c}{TP_c + FP_c + FN_c}) \quad (۸)$$

که در آن c تعداد کل طبقه‌ها است.

۱-۴- نتایج تجربی بر روی مجموعه‌دادهٔ NYUv2

در این بخش، نتایج حاصل از اجرای الگوریتم پیشنهادی بر روی تصاویر RGB-D مجموعه‌دادهٔ NYUv2 [۲۱] با نتایج گزارش‌شده در پنج مقالهٔ مطرح دیگر مقایسه شده‌است.

مجموعه‌دادهٔ استفاده‌شده در این مقاله شامل ۱۴۴۹ تصویر RGB-D از ۸۹۴ طبقه مختلف شی است. هر تصویر براساس طبقه اشیا و عمق آن‌ها در یک صحنه برچسب‌گذاری شده‌اند. در این مجموعه درصورت وجود چندین نمونه از یک طبقه شیء در یک تصویر، هر نمونه یک عدد نمونهٔ منحصربه‌فرد دریافت می‌کند؛ برای مثال، دو صندلی در یک صحنه به‌عنوان دو شی متفاوت برچسب‌گذاری می‌شوند تا به‌طور منحصربه‌فرد آن‌ها را شناسایی کنند.

(جدول-۳): دقت روش‌های بخش‌بندی بر روی مجموعه‌دادهٔ NYUv2، بر اساس سه پارامتر PA, mPA, FWIoU را برای روش‌های مختلف توصیف می‌کند؛ هرچه نتیجهٔ بخش‌بندی به داده‌های برچسب‌گذاری‌شده نزدیک‌تر باشد، سه شاخص مقادیر بالاتری را نشان می‌دهند.

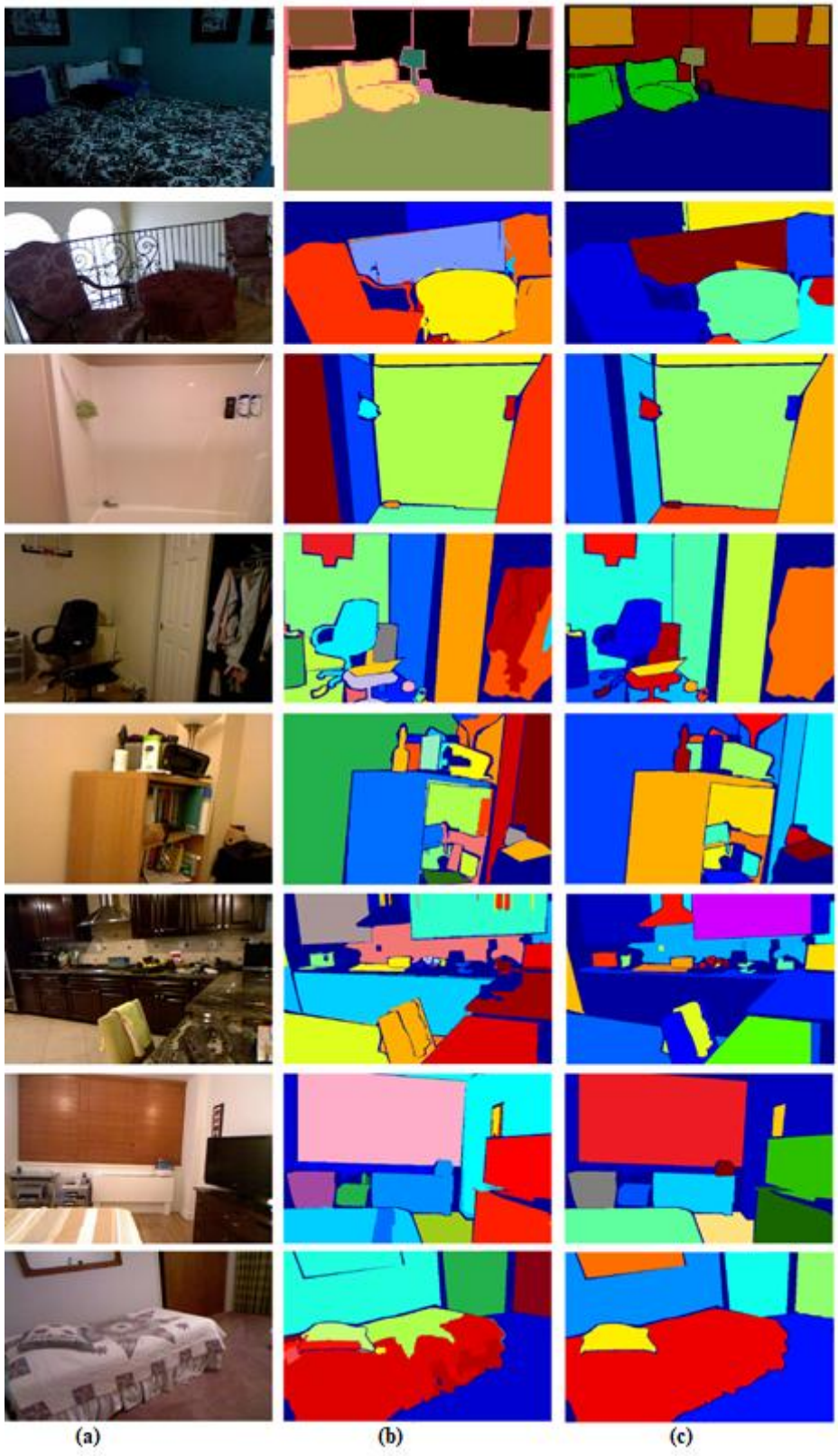
(Table- 3): describes the accuracy of segmentation methods on the NYUv2 dataset based on three parameters: PA, mPA, and FWIoU for various methods. The closer the segmentation result is to the labeled data, the higher the values of these three indicators.

روش	PA	mPA	FWIoU
POR [34]	۵۹.۱	۲۸.۴	۲۹.۱
RGB-D R-CNN[35]	۶۰.۳	۳۵.۱	۳۱.۳
FFCANet[36]	۷۸.۴	۶۵.۳	۵۳.۳
TransD-Fusion[2]	۷۸.۵	۶۹.۴	۵۵.۵
SGACNet[37]	۷۵.۶	۶۲.۷	۴۹.۴
روش پیشنهادی (CRFCut)	۷۸.۷	۶۹.۸	۵۵.۷

در تمام تصاویر این مجموعه‌دادهٔ اشیای ناشناخته با صفر برچسب‌گذاری شده‌ان؛. به‌دلیل آنکه این منبع برای آزمون تعداد زیادی از روش‌های بخش‌بندی معنایی قرار گرفته، محاسبه و تحلیل نتایج شبیه‌سازی این مقاله نیز بر اساس داده‌ها و معیارهای ذکرشده در همین منبع انجام شده‌است؛ همان‌طور که در جدول (۳) نمایش داده شده‌است، در نتایج به‌دست‌آمده، CRFCut از سایر روش‌ها عملکرد بهتری داشته

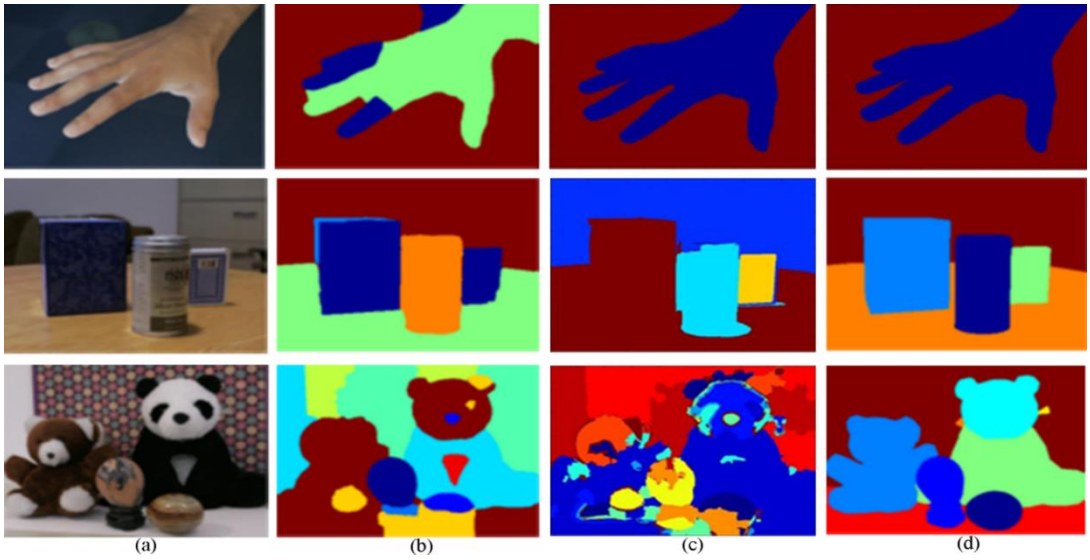
[1] False Positive
[2] False Negative

است. در تحلیل نتایج جدول (۳) مشاهده می‌شود که روش CRFCut با دست‌یابی به PA برابر ۷۸.۷ درصد و mPA برابر ۶۹.۸ درصد، به‌ترتیب ۰.۲ و ۰.۴ واحد درصد بالاتر از بهترین روش پیشین (TransD-Fusion) عمل کرده‌است.



(شکل- ۴): مناطق ارزیابی برای برخی از تصاویر از مجموعه‌دادهٔ NYUv2. در هر ستون (a) تصاویر اصلی، (b) نتیجهٔ بخش‌بندی کد رنگی CRFCut و (c) داده‌های برچسب‌گذاری‌شده توسط انسان.

(Figure- 4): Evaluation regions for some images from the NYUv2 dataset. Each column shows: (a) Original images, (b) Segmentation results with color-coded CRFCut, and (c) Ground truth



(شکل- ۵): نتایج بخش‌بندی در مجموعه‌دادهٔ MIT در هر ستون (a) تصاویر اصلی، (b) نتایج بخش‌بندی nLayers ، (c) نتایج CRFCut (d) داده‌های برچسب‌گذاری‌شده توسط انسان را نشان می‌دهد.

(Figure- 5): Segmentation results on the MIT dataset. Each column shows: (a) Original images, (b) Segmentation results by nLayers, (c) Segmentation results by CRFCut, and (d) Human-labeled data.

(جدول- ۴): اندازه‌گیری RandIndex در تمام نه دنباله از مجموعه‌دادهٔ MIT برای هر چهار روش (و واریانس).

(Table- 4): Rand Index measurements across all 9 sequences from the MIT dataset for each of the four methods (including variance).

	Avg	P-v	ماشین	ماشین ۲	ماشین ۳	سگ	تلفن	میز	اسباب بازی	دست	شخص
HGVS[3]	۰.۵۵۰	۰.۰۰۸	۰.۶۰۲	۰.۴۰۱	۰.۶۸۹	۰.۲۶۰	۰.۴۹۳	۰.۷۶۶	۰.۸۰۹	۰.۴۹۹	۰.۴۳۰
Layer++[5]	۰.۷۷۵	۰.۰۵۰	۰.۶۱۲	۰.۵۱۲	۰.۷۷۸	۰.۹۶۴	۰.۵۶۷	۰.۹۰۹	۰.۸۳۲	۰.۸۱۴	۰.۹۸۶
nLayers[7]	۰.۸۲۳	*	۰.۸۳۶	۰.۵۸۹	۰.۷۶۶	۰.۹۷۴	۰.۵۷۸	۰.۹۷۹	۰.۸۵۸	۰.۸۸۱	۰.۹۴۴
CRFCut	۰.۷۵۴	۰.۰۷۰	۰.۶۵۴	۰.۵۱۰	۰.۷۲۱	۰.۸۱۰	۰.۵۸۶	۰.۸۴۰	۰.۸۶۲	۰.۹۵۸	۰.۹۳۴

افزون‌براین، FWIoU آن با ۵۵.۷ درصد، ۰.۲ درصد بیشتر از TransD-Fusion است؛ این بهبود اندک، اما مستمر از ادغام سلسله‌مراتبی اطلاعات عمق و سوپرپیکسل‌های RGB در مدل CRF، کمینه‌سازی انرژی CRF با الگوریتم برش گراف گسترش آلفا و پردازش بازگشتی موازی به‌دست آمده‌است؛ در واقع، آستانه‌بندی دقیق عمق و تنظیم هوشمند پارامترهای وزن‌های یونری و پی‌وایز CRF ضمن حفظ انسجام نواحی، دقت مرزگشایی را ارتقا داده و باعث شده‌است CRFCut با ساختار بدون نظارت کامل، تمام سه معیار کلیدی ارزیابی را نسبت به بهترین روش گذشته بهبود دهد. شکل (۳) برخی از تصاویر مجموعه‌دادهٔ NYUv2 را با اشیای تقسیم‌شده با دو روش مشخص‌شده با مرزهای قرمز نشان می‌دهد. در این شکل تنها نتایج روش پیشنهادی و بخش‌بندی معنایی انجام‌شده به‌وسیلهٔ الگوریتم [۲] که بهترین عملکرد را بین مقالات بررسی‌شده داشته، آورده شده‌است؛ همان‌طور که به‌وضوح قابل مشاهده است، نتایج بخش‌بندی هر دو روش بسیار عالی است؛ زیرا در بسیاری از مواقع بخش‌بندی به‌درستی انجام شده و در اندک مواردی که برخی از اشیا در عمق و ویژگی‌های ظاهری آن‌ها هم‌زمان تغییر می‌کنند هر دو روش دچار ضعف‌اند که البته الگوریتم پیشنهادی در این مقاله در بیشتر مواقع عملکرد بهتری داشته است؛ زیرا با وجود تکه‌شدن اشیا همچنان شمارهٔ شیء درنتیجهٔ نهایی درست تشخیص داده شده‌است. برای روشن‌شدن چگونگی عملکرد الگوریتم پیشنهادی و مقایسه و بررسی بهتر، نتایج بیشتری از این روش در شکل (۴) نشان داده شده‌است. در این شکل، هر رنگ یک شیء جداگانه را نشان می‌دهد. در تمام تصاویر، مناطقی که با آبی تیره برچسب‌گذاری شده‌اند، مناطق ناشناخته در مجموعه‌داده هستند و باید از نتیجهٔ بخش‌بندی حذف شوند. در بیشتر موارد، روش پیشنهادی به‌خوبی کار می‌کند، اما در برخی موارد نادر مانند ردیف ششم شکل (۳) که دو جسم مجاور در یک سطح عمق قرار می‌گیرند و ویژگی‌های ظاهری آن‌ها نیز مشابه است (برای مثال تلویزیون و میز آن)، روش پیشنهادی به‌طور طبیعی نمی‌تواند آن‌ها را به دو جسم مختلف تقسیم کند و در این تصویر هر دو شی با یک کد رنگ نمایش داده شده‌اند.

۲-۴-نتایج تجربی بر روی مجموعه‌دادهٔ MIT

مجموعه‌دادهٔ MIT در ابتدا برای ارزیابی بخش‌بندی لایه‌ای و تخمین جریان نوری ایجاد شد. مجموعه‌داده شامل نُه دنباله از صحنه‌هایی از داخل اتاق و خارج از اتاق به‌همراه برچسب‌گذاری هر صحنه برای بخش‌بندی لایه‌ای هر قاب توسط انسان است؛ از آنجایی که هدف مقایسهٔ الگوریتم بخش‌بندی لایه‌ای خود با روش‌های دیگر در این مجموعه‌داده بود، ابتدا تصاویر عمقی برای هر قاب از این دنباله‌ها ایجاد شدند. انتخاب پژوهش‌گران در این مقاله روش تخمین عمق پیشنهادشده توسط ژانگ و همکاران بود[۳۸]. با استفاده از مقادیر عمق تخمینی برای هر فاب، می‌توان تصاویر را به لایه‌ها تقسیم کرد. روش اندازه‌گیری دقت در کارهای پیشین، اندازه‌گیری RandIndex بود؛ بنابراین همچنین از معیار RandIndex [۳۹] برای ارزیابی روش خود استفاده می‌شود.

اندازه‌گیری‌های RandIndex در تمام نُه دنباله در جدول (۴) خلاصه شده‌است. در بیشتر مواقع عملکرد ضعیف روش پیشنهادی به‌خاطر اطلاعات نادقیق عمق به‌دست‌آمده به کمک روش‌های محاسبه عمقی بوده که در این مقاله استفاده شده‌است که چون محاسبهٔ عمق هدف این مقاله نبود؛ لذا در محاسبهٔ آن از الگوریتم‌های موجود استفاده شد که آن‌ها نیز به‌دلیل روش به‌کاربرده‌شده جهت تخمین عمق از روی چند تصویر دوبعدی، کارایی لازم برای استفاده در الگوریتم بخش‌بندی را نداشته‌اند؛ بااین‌حال، همان‌طور که در شکل (۵) مشهود است، الگوریتم پیشنهادی عملکرد بسیار قابل قبولی داشته‌است. پژوهش‌گران مقالهٔ حاضر معتقدند که اگر داده‌های عمق دقیق محاسبه شود، روش پیشنهادی در این مقاله می‌تواند از تمام روش‌های دیگری که بررسی شده‌اند بهتر عمل کند. شکل (۵) برخی از نتایج بخش‌بندی روش پیشنهادی را به‌همراه نتایج روش nLayers نشان می‌دهد.

۵- نتیجه‌گیری

در این مقاله یک روش جدید به‌نام CRFCut برای بخش‌بندی لایه‌ای یک صحنه با جفت‌کردن داده‌های رنگ و عمق پیشنهاد شد؛ همچنین یک مدل CRF معرفی شد که می‌توانست پیکسل‌ها را بدون هیچ مجموعهٔ آموزشی برچسب‌گذاری کند. نشان داده شد که توابع انرژی متشکل از مدل CRF را می‌توان با استفاده از الگوریتم بسط آلفا بر اساس برش نمودار به کمترین حد رساند. در ادامه، رویکرد مورد نظر روی دو مجموعه‌دادهٔ بسیار متفاوت آزمایش شد؛ یکی با تصاویر عمق ایجادشده با استفاده از دوربین عمق و دومی با داده‌های عمق تخمین‌زده‌شده با استفاده از نرم‌افزار تخمین عمق. نتایج هر دو آزمایش تأیید کرد که روش پیشنهادی می‌تواند از پیشرفته‌ترین روش‌ها در بسیاری از مواقع عملکرد بهتری داشته باشد. آزمایش‌های انجام‌شده بر روی تصاویر عمقی ایجادشده با نرم‌افزار نشان داد که این روش می‌تواند نتایج بخش‌بندی خوبی را با استفاده از اطلاعات عمق نادقیق حفظ کند؛ هیچ روشی بدون محدودیت نیست و این امر در مورد روش ارائه‌شده در این مقاله نیز صادق است؛ ازجمله محدودیت‌های روش پیشنهادی انتشار خطا در سطوح بالاتر بخش‌بندی به سطوح پایین‌تر است که این امر باعث ایجاد مناطقی بیش از مناطق مورد انتظار در نتایج بخش‌بندی نهایی می‌شود؛ به‌این‌ترتیب، ارائهٔ روشی برای افزودن نواحی بخش‌بندی‌شدهٔ کوچک به منطقهٔ مناسب خود با تعریف مجدد ناحیهٔ مورد نظر با درنظرگرفتن مناطق بیش‌ازحد بخش‌بندی‌شده، کار پژوهشی آیندهٔ نویسندگان مقاله خواهد بود؛ همچنین، مانند سایر روش‌های بخش‌بندی بدون نظارت، روش پیشنهادی تمایز بین دو شیء مجاور مختلف با ویژگی‌های ظاهری مشابه در سطح عمق یکسان را به‌درستی تشخیص نمی‌دهد. در آینده ممکن است، بتوان این مشکل را با استفاده از مقدار زیادی نمونه برای هر شیء در یک مجموعهٔ آموزشی و ترکیب روش پیشنهادی با یک روش نظارت‌شده حل کرد.

۶-مراجع 6-Reference

[1] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 24, no. 5, pp. 603-619, 2002.

[2] Z. Wu, Z. Zhou, G. Allibert, C. Stolz, C. Demonceaux, and C .Ma, "Transformer fusion for indoor rgb-d semantic segmentation," *Computer Vision and Image Understanding,* vol. 249, p. 104174, 2024.

[3] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss, "Human-assisted motion annotation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008: IEEE, pp. 1-8 .

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision,* vol. 59, pp. 167-181, 2004.

[5] D. Sun, E. B. Sudderth, and M. J. Black, "Layered segmentation and optical flow estimation over time," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 1768-1775 .

[6] L. u. Ladický, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical crfs for object class image segmentation," in *2009 IEEE 12th international conference on computer vision*, 2009: IEEE, pp. 739-746 .

[7] Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern*

Recognition (CVPR'06), 2006, vol. 1: IEEE, pp. 53-60 .

[8] M. Szummer, P. Kohli, and D. Hoiem, "Learning CRFs using graph cuts," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, 2008: Springer, pp. 582-595 .

[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23 ,no. 11, pp. 1222-1239, 2001.

[10] C. Rother, V. Kolmogorov, and A. Blake, "" GrabCut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG),* vol. 23, no. 3, pp. 309-314, 2004.

[11] S. Mirkamali and P. Nagabhushan, "Depth-wise image inpainting," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012: IEEE, pp. 141-144 .

[۱۲] حاجی اسماعیلی، محمدمهدی، منتظر، غلامعلی، «مروری نقادانه بر روش‌های بازیابی محتوامحور و معناگرای تصاویر»، *فصلنامهٔ پردازش علائم و داده‌ها*، ۲۲ (۱)، صص ۱۱۳-۱۴۱، ۱۴۰۴.

[12] M. M. Haji-Esmaeili and G. Montazer, "a Critical Survey on Content-Based & Semantic Image Retrieval – Abstract," (in eng), *Signal and Data Processing,* Research vol. 22, no. 1, pp. 113-141, 2025, doi: 10.61186/jsdp.22.1.113.

[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, no. 8, pp. 888-905, 2000.

[14] S. Du, W. Wang, R. Guo, R. Wang, and S. Tang, "Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7608-7615 .

[15] X. He, R. S. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 2006,7-13 Proceedings, Part I 9*, 2006: Springer, pp. 338-351 .

[16] Ren and Malik, "Learning a classification model for segmentation," in *Proceedings ninth IEEE international conference on computer vision*, 2003: IEEE, pp. 10-17 vol. 1 .

[17] A. Jepson and M. J. Black, "Mixture models for optical flow computation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1993: IEEE, pp. 760-761 .

[18] N. Jojic and B. J. Frey, "Learning flexible sprites in video layers," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1: IEEE, pp. I-I .

[19] D. Sun, E. Sudderth, and M. Black, "Layered image motion with explicit occlusions, temporal consistency, and depth ordering," *Advances in Neural Information Processing Systems,* vol. 23, 2010.

[20] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo—joint stereo matching and object segmentation," in *CVPR 2011*, 2011: IEEE, pp. 3081-3088 .

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 2012: Springer, pp. 746-760 .

[22] L. Wang, C. Zhang, R. Yang, and C. Zhang, "Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera," in *Proc. of 3DPVT*, 2010, pp. 1-8 .

[23] A. D. Jepson, D. J. Fleet, and M. J. Black, "A layered motion representation with occlusion and compact spatial support," in *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*, 2002: Springer, pp. 692-706 .

[23] Y. Weiss and E. H. Adelson" ,A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models," in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996: IEEE, pp. 321-326 .

[25] J .Wills, Agarwal, S., and Belongie, S., "What Went Where," *CVPR,* vol. v.1, pp. 37-44, 2003.

[26] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *IEEE transactions on pattern analysis and machine intelligence,* vol. 27, no. 10, pp. 1644-1659, 2005.

[27] P. Kohli, L. u. Ladický, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision,* vol. 82, pp. 302-324, 2009.

[28] B. Yin, X. Zhang, Z. Li, L. Liu ,M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," *arXiv preprint arXiv:2309.09668,* 2023.

[29] L. Zhong, C. Guo, J. Zhan, and J. Deng, "Attention-based fusion network for RGB-D semantic segmentation," *Neurocomputing,* vol. 608, p. 128371, 2024.

[30] Z. Li, C. Lang, G. Li, T. Wang, and Y. Li, "Depth guided feature selection for RGBD salient object detection," *Neurocomputing,* vol. 519, pp. 57-68, 2023.

[31] Y. Tong, J. Chen, and Y. Wang, "Geometry-guided multilevel RGBD fusion for surface normal estimation," *Computer Communications,* vol. 206, pp. 73-84, 2023.

[32] B. Xiong, Y. Peng, J. Zhu, J. Gu, Z. Chen, and W. Qin, "AGWNet: Attention-guided adaptive shuffle channel gate warped feature

network for indoor scene RGB-D semantic segmentation," *Displays,* p. 102730, 2024.

[33] N. Komodakis, G. Tziritas, and N. Paragios, "Fast, approximately optimal solutions for single and dynamic MRFs," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007 :IEEE, pp. 1-8 .

[34] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 564-571 .

[35] S. Gupta, R .Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, 2014: Springer, pp. 345-360 .

[36] Y. Liu, O. Yoshie, and H. Watanabe, "Application of multi-modal fusion attention mechanism in semantic segmentation," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 1245-1264 .

[37] Y. Zhang, C. Xiong, J. Liu, X. Ye, and G. Sun, "Spatial-information guided adaptive context-aware network for efficient RGB-D semantic segmentation," *IEEE Sensors Journal,* 2023.

[38] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, "Consistent depth maps recovery from a video sequence "*,IEEE Transactions on pattern analysis and machine intelligence,* vol. 31, no. 6, pp. 974-988, 2009.

[39] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association,* vol. 66, no. 336, pp. 846-850, 1971.

دکتر سیدسعید میرکمالی رئیس دانشگاه پیام‌نور مرکز بهشهر و استادیار گروه مهندسی و فناوری اطلاعات کامپیوتر دانشگاه پیام‌نور است. وی دانش‌آموختهٔ دکترای کامپیوتر گرایش پردازش تصویر و کارشناسی‌ارشد رشتهٔ مهندسی هوش مصنوعی از دانشگاه میسور هند است. از تألیفات ایشان می‌توان به چاپ دو کتاب و بیش از بیست مقالهٔ علمی و پژوهشی اشاره کرد؛ وی همچنین در حال حاضر با بیش از ده مجله و کنفرانس بین‌المللی به‌عنوان داور و دبیر علمی همکاری مستمر دارد؛ همچنین ایشان بنیان‌گذار و مشاور چند شتاب‌دهنده در حوزه‌های فناوری اطلاعات و هوش مصنوعی نیز بوده‌است.

نشانی رایانامهٔ ایشان عبارت است از:

s.mirkamali@pnu.ac.ir