با نام خدا

درس یادگیری عمیق

پروژه : تشخیص عمق اشیاء

سروش نورزاد (99205372) - ارشد الکترونیک دیجیتال

ابوالفضل فلاح پاکدامن (99205326) - ارشد الکترونیک دیجیتال

بهمن ماه 1400

دانشگاه صنعتی شریف

نکات عمومی در خصوص پروژه :

- جواب تمارین شماره یک و دو و سه کامپیوتری در قالب سه نوت بوک q1-3 تحویل گردیده است.
- نسخه ی 2.7 از کتابخانه ی تنسورفلو مورد استفاده قرار گرفته است.
- به علت کمبود وقت، گزارش به تلخیص نوشته شده است.
- ممکن است به هنگام pdf کردن فایل ورد، کیفیت تصاویر پایین آمده باشد. در این صورت می توانید تصاویر اصلی را در فایل های ipynb ارسالی در کنار کدهای اجرا شده مشاهده نمایید.
- به علت طولانی بودن تمارین، تا حد ممکن از پردازش های زمان گیر خودداری کرده و با تعداد ایپاک های کمتری آزمایشات انجام شده که نتیجه ی آن پایین تر بودن کیفیت خروجی هاست. اما نتایج به طور کلی گویای خروجی مطلوب هستند.

توضیحات ابتدایی:

توضیحات عمومی دیتاست NYU-Depth-V2:

این دیتاست که از تصاویری که از محیط های بسته تصویر برداری شده است تشکیل شده است. این تصاویر با استفاده از کینکت مایکروسافت که توانایی تصویر برداری سه بعدی را داراست تصویر برداری شده است. کینکت می تواند هم تصویر RGB و هم تصویر نمایش دهنده ی عمق را ذخیره کند. این دیتاست 1449 جفت تصویر RGBD (تصویر RGB و تصویر نمایش دهنده ی عمق آن D) از 464 صحنه ی متفاوت از سه شهر متفاوت را دارا می باشد. در این دیتاست، از روش لیبل گذاری بر پیکسل بهره گرفته شده است که با کمک سامانه ی خدماتی Amazon Mechanical Turk انجام گرفته است. هر کدام از اشیاء تصویر با یک کلاس و شماره ی نمونه (instance) آن لیبل گذاری شده است. نحوه ی شماره گذاری هر نمونه به این شکل است که اگر در هر تصویری، چند نمونه از یک کلاس موجود باشد، شماره نمونه های متفاوتی به نمونه ها تخصیص داده می شود.

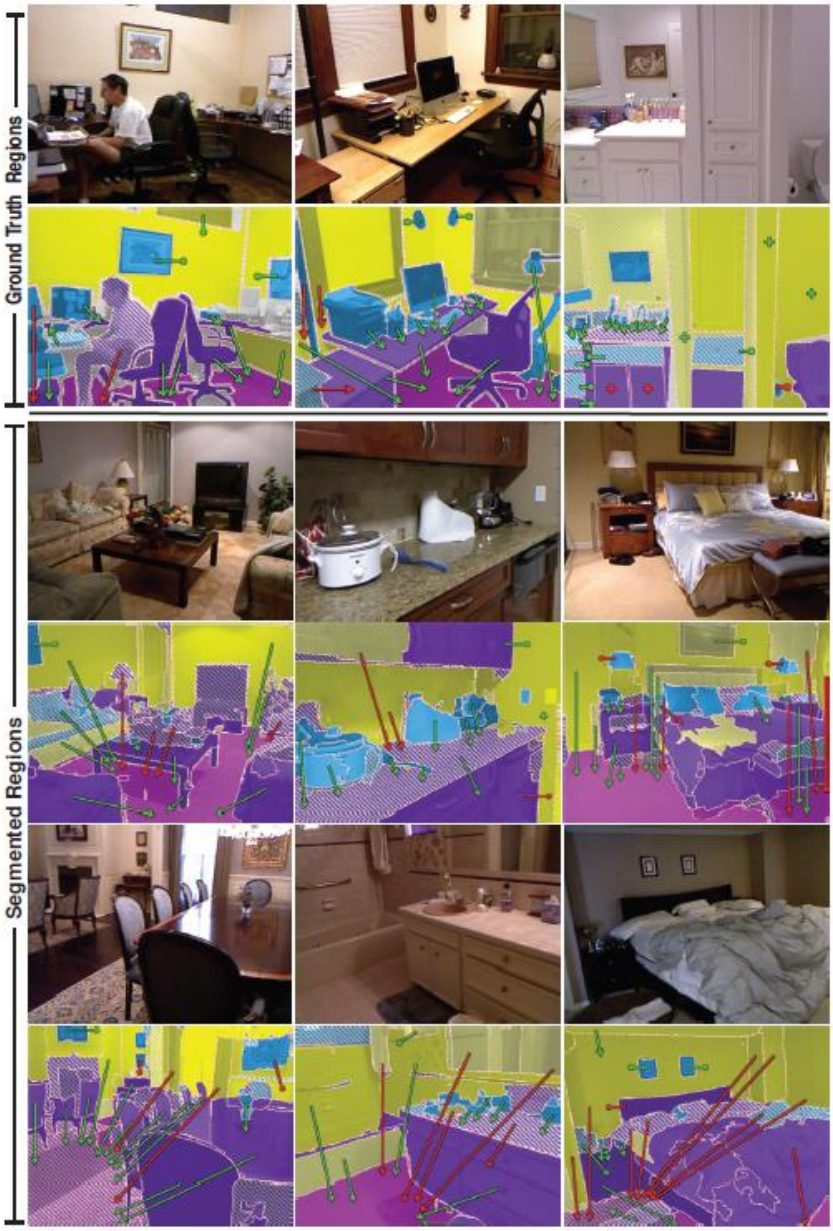


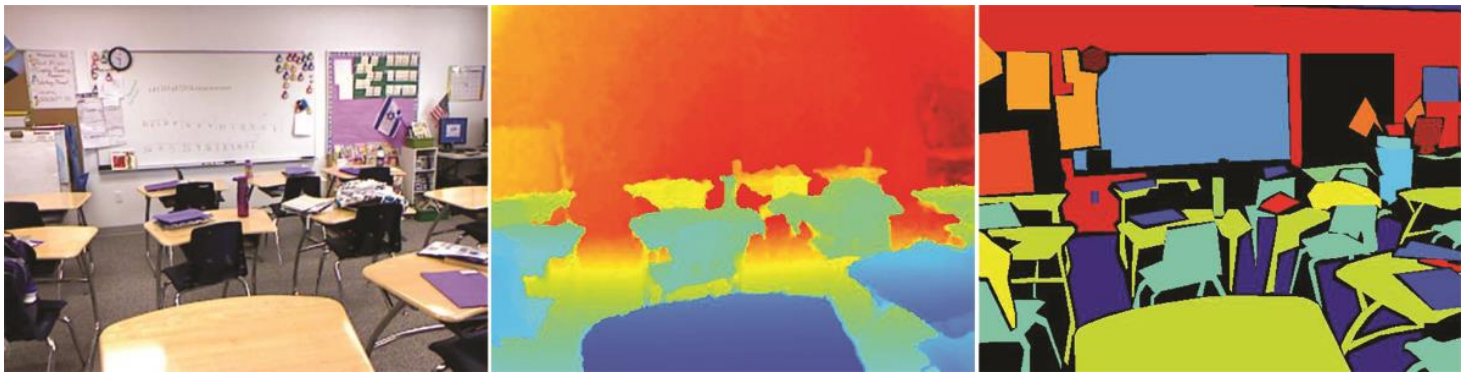
Fig. 7. Examples of support and structure class inference with the LP solution. → : support from below, —o : support from behind, + : support from hidden region. Correct support predictions in green, incorrect in red. Ground in pink, Furniture in Purple, Props in Blue, Structure in Yellow, Grey indicates missing structure class label. Incorrect structure predictions are striped.

دیتاست شامل 36064 شی مجزا از یکدیگر است که در 894 دسته ی متفاوت (label) جای گرفته اند. برای تمام 1449 تصویر موجود در دیتابیس بردار اتصال (Support annotations) به طور دستی اضافه شده است. هر بردار اتصال شامل یک مجموعه ی سه تایی از Ri (شماره (ID) ناحیه (Region) ی شی حمایت شده) ، Rj (شماره (ID) ناحیه (Region) ی شی حمایت کننده) و نوع بردار است. بردار انواع متفاوتی را داراست که به طور مثال یکی حمایت از پایین به بالا (حمایت شده روی حمایت کننده قرار می گیرد. مانند فنجان روی میز) و یکی حمایت از پشت به جلو است (حمایت کننده از نقطه ای دورتر به حمایت شونده که نزدیک تر است وصل شده است، مانند قاب روی دیوار). در تصویر روبرو این موضوع قابل مشاهده است.

مجموعه فایل های ارائه شده شامل چند بخش است:

- دیتاست Labeled: یک زیر مجموعه از تمامی داده های موجود است که با لیبل های چند کلاسه و سایر اطلاعاتش همراه شده است. در این داده ها پیش پردازش های لازم نیز به منظور امکان کار کردن با داده ها صورت پذیرفته است.
- فایل های Raw: تصاویر خام RGB و تصاویر Depth و Accelerometer متناظر با تصویر RGB که توسط Kinect مایکروسافت ضبط شده اند.
- فایل های Toolbox: ابزارهای لازم به جهت کار کردن با داده ها در متلب.

ما در این پروژه از دیتاست دارای لیبل بهره خواهیم گرفت. چرا که برای آموزش مدل ها نیاز به آموزش با ناظر داریم. در این دیتاست، تصاویر و اطلاعات آنها به صورت داده هایی ذخیره شده اند. تصاویر ذخیره شده، شامل خود تصاویر RGB ، تصاویر دارای مشخصه ی عمق تصویر و تصاویر دارای کلاسه بندی می باشند. نمونه ای از این تصاویر را می توان در شکل زیر مشاهده کرد:



دیتاست ذخیره شده، شامل موارد زیر است (N تعداد داده ها، H ارتفاع تصاویر و W عرض تصاویر محسوب می شوند):

- accelData: ماتریسی N*4 از مقادیر شتاب در زمان گرفته شدن تصاویر. هر کدام از N مقدار این داده ها شامل چهار پارامتر غلتش (roll) ، نوسان (yaw) ، گام (pitch) و زاویه ی کج شدن (tilt angle) دستگاه می باشند.
- depths: ماتریسی N*H*W که نمایش دهنده عمق N تصویر موجود در دیتاست می باشند.
- Images: ماتریسیN*3*H*W که همانطور که مشخص است، سه کانال RGB تصاویر اصلی را در بر دارد.
- instances: ماتریسی N*H*W که مختصه های نمونه های موجود در تصاویر را به همراه دارد. می توان از ابزار get_instance_masks.m نوشته شده در متلب برای بیرون کشیدن داده های موجود در این بخش استفاده کرد.
- labels: ماتریسی N*H*W که ماسک های لیبل ها را نمایش می دهند. لیبل ها از 1 تا C هستند که C تعداد کل کلاس های موجود می باشند. لیبل صفر به منزله ی لیبل نخورده بودنِ پیکسل است.
- names: آرایه ای C تایی که نام انگلیسی تمام لیبل ها به ترتیب در آن قرار دارد.
- namesTolds: نگاشتی از نام های انگلیسی کلاس ها به ID کلاس ها (تعداد C جفتِ 2 تایی key ، value).
- rawDepths: ماتریسی N*H*W که نمایشگر نگاشت "تصاویر خام نمایشگرِ عمق" هستند. این نگاشت ها عمق تصاویر را مانند بخش depths نشان می دهند، با این تفاوت که در اینجا، بخش های از دست رفته و تشخیص داده نشده، بهبود پیدا نکرده اند و پیش پردازش های لازم روی آنها انجام نگرفته است. به طور مثال هنوز غیرخطی بودن تصاویر دریافتی از کینکت در این داده ها، حذف نشده است.

- rawDepthFileNames: ماتریسی N*1 که نام تصاویر موجود در دیتاست Raw که به منظور تولید داده ی متناظر از آن بهره گرفته شده است را دارا می باشد.
- rawRgbFileNames: ماتریسی N*1 که نام تصاویر اصلی موجود در دیتاست Raw را دارا می باشد.
- scenes: ماتریسی N*1 که نام صحنه ای که تصویر از آن گرفته شده است را دارا می باشد.
- sceneTypes: ماترسی N*1 که نوع صحنه ای که تصویر از آن گرفته شده است را دارا می باشد.

بخش Toolbox (شاید نیاز به حذف این بخش باشد، چون نیاز به توضیح این بخش نیست. اما در عمل به تحلیل توابع این بخش ممکن است نیاز داشته باشیم):

The matlab toolbox has several useful functions for handling the data.

- `camera_params.m` – Contains the camera parameters for the Kinect used to capture the data.
- `crop_image.m` – Crops an image to use only the area when the depth signal is projected.
- `fill_depth_colorization.m` – Fills in the depth using Levin et al's Colorization method.
- `fill_depth_cross_bf.m` – Fills in the depth using a cross-bilateral filter at multiple scales.
- `get_accel_data.m` – Returns the accelerometer parameters at a specific moment in time.
- `get_instance_masks.m` – Returns a set of binary masks, one for each object instance in an image.
- `get_rgb_depth_overlay.m` – Returns a visualization of the RGB and Depth alignment.
- `get_synched_frames.m` – Returns a set of synchronized RGB and Depth frames that can be used to produced RGBD videos of each scene.
- `get_timestamp_from_filename.m` – Returns the timestamp from the raw dataset filenames. This is useful for sampling the RAW video dumps at even intervals in time.
- `project_depth_map.m` – Projects the Depth map from the Kinect on the RGB image plane.

در روش پیاده سازی شده ی مقاله که در تصویر بعدی نیز قابل مشاهده است، ابتدا ساختار تصویر سه بعدی مورد نظر را پردازش کرده و سپس تصویر را به اشیا مجزایی که در ارتباط با یکدیگر هستند تقسیم کرده و ارتباط آنها با یکدیگر مشخص شده است. کشف بعضی از این ارتباطات ، مانند جهت و عمق قرارگیری سقف و دیوار ها، آسان و کشف بعضی مانند تشخیص هر شی و قطعه بندی تصویر می تواند کار سختی باشد. بنابراین از نگرش عمق محور (Depth cues) به منظور مرتفع کردن چالش های هندسی بهره گرفته شده است که باعث می شود تصویر تبدیل به ساختاری با جزئیات بیشتر شود. پس از انجام این عملیات و به دست آوردن ساختار عمق محور، می توان با سهولت بیشتری، اجزای تصویر را از یکدیگر جدا ساخته و رابطه ی اجزا با یکدیگر را نیز بهتر شناسایی کرد. یکی از ابتکارات موجود در مقاله، دسته بندی کلاس ها به کلاس های ساختاری (Structural Classes) است. با استفاده از این روش می توان نقش فیزیکی هر عنصر در تصویر را به خوبی شناسایی کرد. دسته های ساختاری را می توان شامل مواردی چون زمین (Ground)، ساختارهای دائمی (Permanent structures) مانند دیوار ها، سقف ها و ستون ها، وسایل بزرگ (large furniture)، مانند میزها، پیشخوان ها و کمدها و اثاثیه (props) که قابل حرکت دادن هستند، دانست.

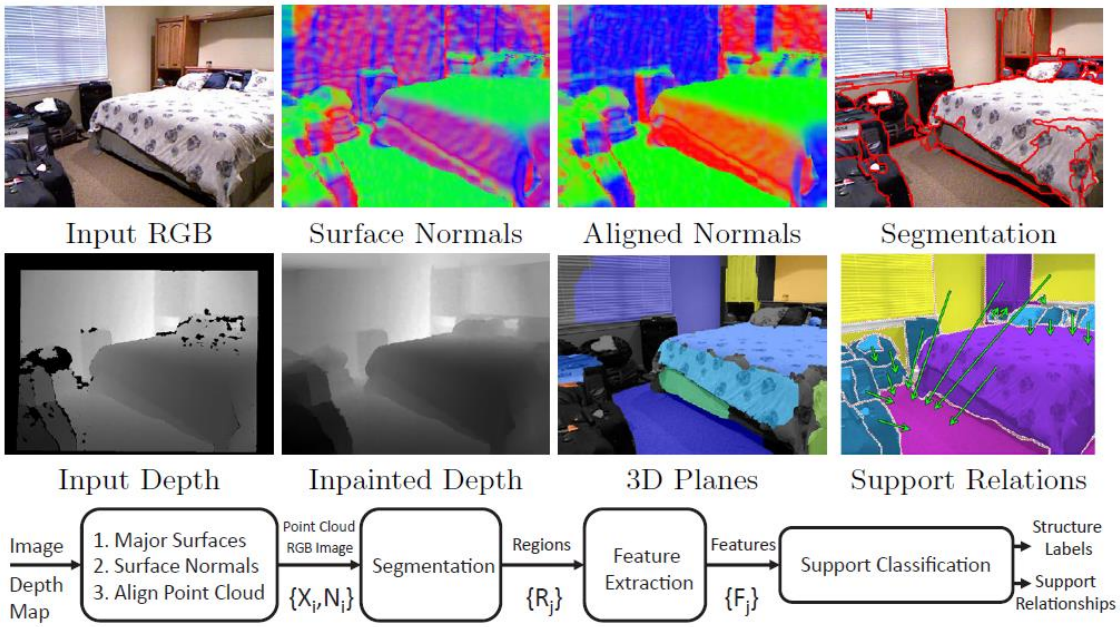


Fig. 1. Overview of algorithm. Our algorithm flows from left to right. Given an input image with raw and inpainted depth maps, we compute surface normals and align them to the room by finding three dominant orthogonal directions. We then fit planes to the points using RANSAC and segment them based on depth and color gradients. Given the 3D scene structure and initial estimates of physical support, we then create a hierarchical segmentation and infer the support structure. In the surface normal images, the absolute value of the three normal directions is stored in the R, G, and B channels. The 3D planes are indicated by separate colors. Segmentation is indicated by red boundaries. Arrows point from the supported object to the surface that supports it.

برای آنکه بتوان رابطه های بخش های متفاوت با یکدیگر را شناسایی کرد، روشی ارائه شده است که قیدهای فیزیکی را با مقدمات آماری ممکن (رابطه های غالب اشیا در جهان پیرامون) پیوند بزند. این روش روی تصاویر شلوغ و واقعی با المان های بسیار زیاد نیز جواب داده است. در این گونه ی تصاویر، رابطه های موجود بین بخش های متفاوت در تصویر به خوبی قابل مشاهده نیست و باید به گونه ای استنتاج شود. از دیگر مشکلات موجود در تصاویر واقعی، تغییرات قابل توجه در عمق های موجود در تصویر است. علاوه بر مشکلاتی که در تصاویری با موقعیت های مکانی گسترده دارند موجود است، در تصاویری که به بخشی از یک محیط کوچک نیز اشاره می کنند، ممکن است به چالش هایی نظیر عدم توانایی در تشخیص کف یا سقف محیط نیز بر بخوریم که با روش موجود در مقاله این چالش نیز بر طرف شده است.

مدل سازی تصاویر داخل ساختمانی:

از آنجایی که فضای درون اتاق ها معمولا دارای خط کشی های عمودی-افقی زیادی هستند و به میزان زیادی، شیار های اینگونه در تصویر وجود دارد، یکی از اولین کارهایی که در عملیات انجام داده اند، تبدیل عملیات به یک مسئله ی alignment و segmentation بوده است. برای این کار ابتدا بردار نرمال های سطوح را با استفاده از مشاهدات موجود در تصویر عمق به دست آورده اند و سپس با استفاده از آنها و همچنین با توجه به خطوط مستیم موجود در تصویر ، شاخص ترین سه جهت اصلی تصویر را محاسبه کرده اند و پس از آن مختصات سه بعدی را به گونه ای در نظر گرفته اند که در تناسب با جهت های اصلی به دست آمده قرار گیرد. سپس صفحات سه بعدی را با استفاده از RANSAC به مختصات سه بعدی تبدیل کرده اند و نهایتا بخش های قابل مشاهده ی تصویر را که نمایانگر این صفحات اصلی می باشند را به عنوان یکی از آن صفحات سه بعدی یا پس زمینه ی تصویر در نظر گرفته اند (با استفاده از Graph cuts روی بردارهای نرمال، نقاط سه بعدی و گرادیان های RGB). در تصویر زیر می توان چند مثال از این عملیات را مشاهده کرد.

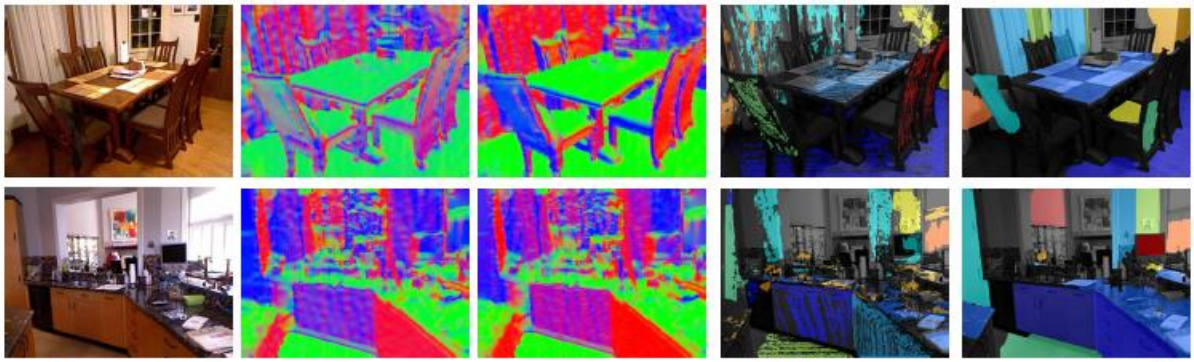


Fig. 2. Scene Structure Estimation. Given an input image (a), we compute surface normals (b) and align the normals (c) to the room. We then use RANSAC to generate several plane candidates which are sorted by number of inliers (d). Finally, we segment the visible portions of the planes using graph cuts (e). Top row: a typical indoor scene with a rectangular layout. Bottom row: an scene with many oblique angles; floor orientation is correctly recovered.

تشخیص نمای مشترک و تخمین عمق در تصویر چندگانه

بهره مندی از توسعه شبکه عصبی پیچشی عمیق، تشخیص شئ، از یک پیشرفت بزرگ در سال های اخیر ساخته شده است. با توجه به تصویر داده شده، هدف تشخیص شئ، برای به دست آوردن مکان و دسته ای از اطلاعات هر نمونه شئ در آن است. به عنوان بخش مهمی از بینایی کامپیوتری، تشخیص اشیاء، طیف وسیعی از برنامه های کاربردی در بسیاری از زمینه ها مانند رانندگی خودمختار، بینایی ربات و سیستم نظارت دارد.

با این حال، برخی از برنامه ها (مانند رانندگی خودکار) نه تنها به موقعیت اشیاء در تصویر نیاز دارند، بلکه همچنین به عمق واقعی این اشیاء شناسایی شده نیز نیاز دارد. این کار را می‌توان با تشخیص شئ سه بعدی که مکان، ابعاد (ارتفاع، عرض و طول) سه بعدی را پیش بینی می‌کند و جهت گیری اشیاء تکمیل کرد. بهره مندی از عمق دقیق اندازه گیری روش های 13 و 15 مبتنی بر داده های (لیدار) (تشخیص نور و محدوده)، عملکرد پیشرفته ای در تشخیص اشیاء سه بعدی به دست میابند. اما (لیدار) دارای معایب هزینه بالا، دامنه درک نسبتاً کوتاه و اطلاعات پراکنده است.

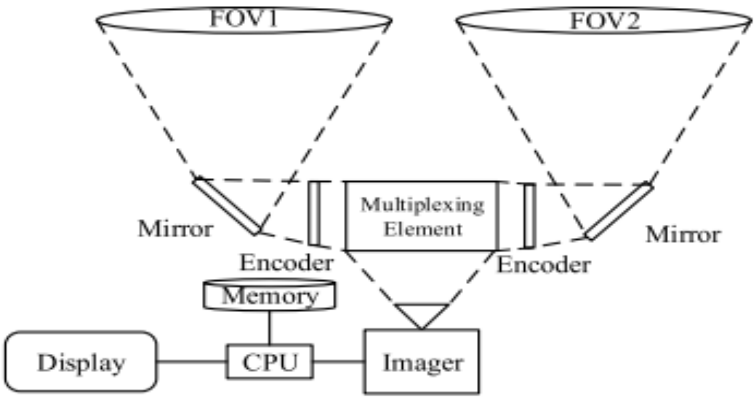
از سوی دیگر روش های 16 و 18، که ورودی آن یک تصویر تک چشمی است، نمی تواند عمق دقیق اجسام را، به خصوص برای صحنه های دیده نشده پیش بینی کند.

استرو (آر سی ان ان) 19، یک روش تشخیص شئ سه بعدی است که از اطلاعات پراکنده و متراکم معنایی و هندسی در تصاویر استریو استفاده می‌کند اما سرعت استنتاج آن (0.28 ثانیه در هر تصویر) از زمان واقعی مورد نیاز برای رانندگی خود مختار، به خاطر استخراج ویژگی از دو تصویر، یک تصویر استریو ویک تصویر پس پردازش، بسیار زیاد است.

برای این منظور، ما یک آشکارساز ساده و سریع را پیشنهاد می‌کنیم که شامل تخمین عمق، بر اساس تصویر چندگانه است.تخمین عمق بر اساس تصویر برداری چندگانه نوری یک زمینه در حال توسعه درحوزه تصویربرداری مخابراتی است.

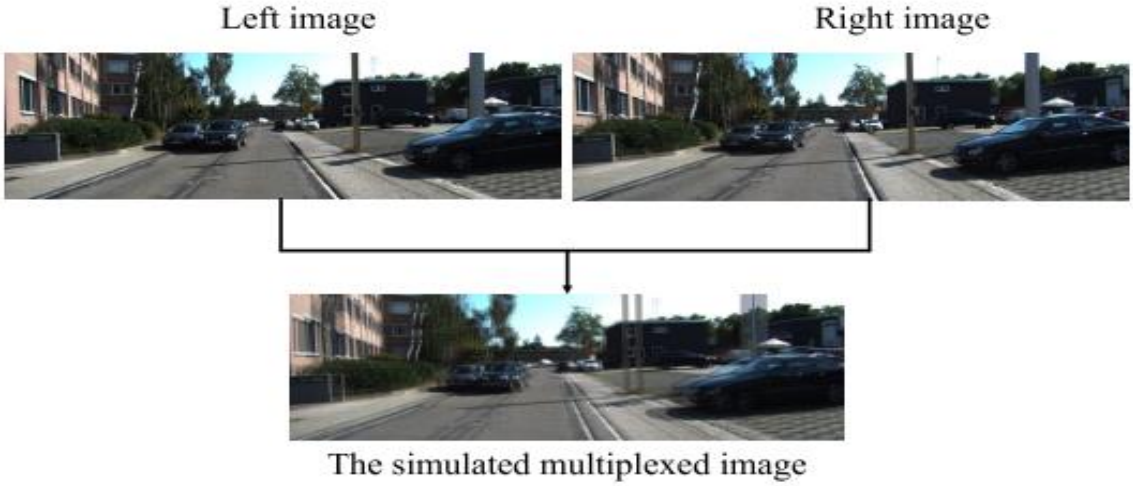
شپرد و راچلین [1] دستگاه تصویربرداری جدید که چندین کانال نور رابه طور همزمان توسط یک سنسور جمع آوری می کند را پیشنهاد کردند.

همانطور که در شکل 1 نشان داده شده است.



شکل 1. معماری دستگاه برای تصویربرداری چندگانه

تصویر سمت چپ تصویر سمت راست



تصویرچندگانه شبیه سازی شده

شکل 2. یک تصویر استریو. تصویر از دوربین سمت چپ و تصویر از دوربین سمت راست و تصویر چندگانه. در این مقاله ما از تصویر همپوشانی برای شبیه سازی تصویر چندگانه استفاده کرده ایم.

آنها همچنین روش‌هایی را برای ابهام زدایی از یک تصویر چندگانه گرفته شده برای ایجاد تصاویر برای هر یک از چندین تصویر در کانال های تصویری که می توانند تصاویر استریو تولید کنند را پیشنهاد کردند. در مقایسه با تصویربرداری چندگانه، روشی که از دو یا چند مورد دوربین برای ایجاد یک تصویر استریو استفاده می کند که از هزینه

های اضافی، قدرت، حجم و پیچیدگی و استفاده از دوربین های متعدد، زیان میبیند. علاوه بر این، روش ما مستقیماً از تصاویر چندگانه به عنوان ورودی به جای تصاویر استریو بازیابی شده از آنها استفاده میکند. این باعث میشود که روش ما به اندازه سرعت یک آشکارساز معمولی باشد.

توجه داشته باشید که هدف از کار ما تخمین عمق است. اطلاعات اشیاء شناسایی شده تصویربرداری چندگانه دراین مقاله دارای دو لنز دوربین افقی مانند تصویربرداری استریو است اما فقط از یک سنسور تصویربرداری استفاده می‌کند. این باعث می شود که تصویرچندگانه با همپوشانی یک تصویر استریو برابری کند. جفت، همانطور که در شکل 2 نشان داده شده است.

کار ما بر اساس مشاهده است که هم ظاهر و هم نابرابری هر شی به طور ضمنی در تصویر چندگانه کدگذاری شده است. روش ما، به نام نابرابری آشکار ساز، ابتدا تمام نماهای یک شئ را به عنوان یک گروه با استفاده ازاستراتژی پیشنهادی ما شناسایی میکند سپس از تفاوت نماهای مختلف، برای تخمین عمق هر جسم استفاده میکند.

هنوز مجموعه داده تصویر چندگانه عمومی وجود ندارد. بنابراین، آزمایش‌ها بر روی تصویرچندگانه شبیه‌سازی شده با استفاده از تصاویر استریو از (کی آی تی تی آی) انجام می‌شوند. روش پیشنهاد شده، بر اساس استقامت (وی جی جی 16) و چهارچوب آشکارساز (اس اس دی) توسعه یافت. اما میتوان آن را به راحتی، با دیگر آشکارساز های شبکه عصبی پیچشی، مبتنی بر لنگر(دی اس اس دی) و استقامت (رس نت) برای عملکرد بهتر، گنجاند. کل خط لوله کار ما در شکل 3 نشان داده شده است. آثار این مقاله در ادامه جالب توجه است: جنبه ها: 1) روش پیشنهادی می تواند به طور همزمان موقعیت های دو بعدی و عمق واقعی اشیاء را در تصویر چندگانه تخمین بزند. 2) مقایسه با شئ محبوب آشکارسازها، روش پیشنهادی تقریباً هیچ بار محاسباتی اضافی یا زمان تأخیر ندارد. 3) حتی در تصاویر چندگانه ترکیبی، روش پیشنهادی به تشخیص رقابتی و نتایج تخمین عمیق دست میابد.

کار مرتبط

در این بخش قصد داریم به طور مختصر به بررسی پیشرفت‌های آثار مرتبط از سه جنبه بپردازیم:

تصویربرداری چندگانه، تشخیص شئ و چارچوب تخمین نابرابری. الف. تصویربرداری چندگانه:

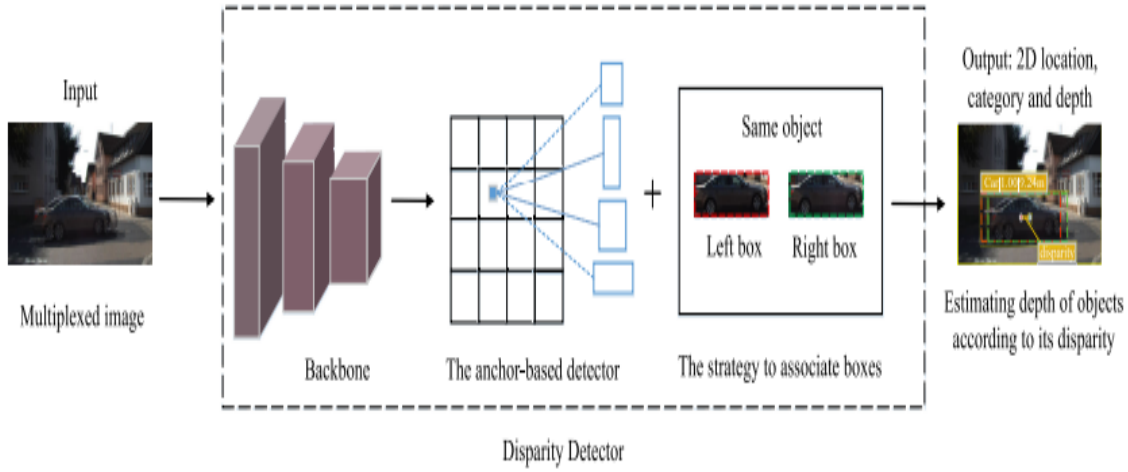
تکنیک های اکتساب تصویر چندگانه نوری، به طور فزاینده ای برای رمزگذاری های مختلف، نوردهی ها، کانال های رنگی، میدان های نوری و سایر ویژگی ها محبوب شده اند. وتسشتاین و همکاران، مدل سازی نوری مشترک و رویکرد بازسازی محاسباتی برای افزایش دامنه دینامیکی عکس های چندگانه ارایه کرده اند. شپرد و راچلین دستگاه‌های تصویربرداری چندگانه و روش‌هایی که تصویرچندگانه گرفته شده برای ایجاد کانال های تصویررفع می‌کنند را پیشنهاد دادند. اوتام و همکاران، یک رده از تصویر سازهای چندگانه مخصوص کار برای جمع آوری داده های رمزگذاری شده در اندازه گیری با ابعاد پایین ترپیشنهاد کرد که نامش فضای انطباق بود و یک الگوریتم رمزگشایی ایجاد کرد که اهداف را مستقیما در این فضای انطباق ردیابی میکند.

ب. تشخیص شئ

هدف از تشخیص اشیا به دست آوردن مکان واطلاعات دسته بندی هر نمونه شئ در یک تصویر معین است.
آشکارساز های کلاسیک ویژگی های هر کشویی را توسط توصیف گر های مهندسی شده دستی استخراج می کنند و سپس

طبقه بندی کننده ها را برای پیدا کردن اشیاء اعمال میکنند. در سالهای اخیر شبکه عصبی پیچشی عمیق به طور گسترده برای وظایف بینایی استفاده میشوند. جدای از آشکار سازهای کلاسیک، آشکار سازهای مبتنی بر شبکه عصبی پیچشی، از تصویر برای ویژگی های استخراج شده توسط یک شبکه پایه به عنوان مثال (وی جی جی) برای یافتن اشیا استفاده میکنند.

با توجه به عملکرد برجسته مبتنی بر شبکه عصبی پیچشی، آشکار سازهای جسم به نیروی اصلی در میدان تشخیص، تبدیل میشوند. معمولا آشکارسازهای مبتنی بر شبکه عصبی پیچشی را میتوان تقریبا به دو دسته تقسیم کرد، دو دسته یعنی رویکرد دو مرحله ای و رویکرد یک مرحله ای، رویکرد دو مرحله ای به عنوان مثال، (آر سی ان ان) و (فست آر سی ان ان) و (فستر آر سی ان ان) دو مرحله دارند که در آن اولی تعداد ثابتی از طرح های شئ بالقوه را تولید میکند و دومی انحراف های فضایی برچسب های مکان و دسته را پیش بینی میکند. روش های دو مرحله ای در چندین معیاراز جمله ام اس و پاسکال ووک و کوکو، به نتایج برتر دست یافته است. اخیرا تعداد زیادی از تکنیک های جدید برای عملکرد بهتر استفاده میشود مانند جعبه محدود کننده تکراری، رگرسیون، استراتژی آموزشی و ضرر جدید برای رگرسیون جعبه مرزی.



شکل 3. کل خط لوله چارچوب آشکارساز نابرابری ما. آشکارساز اختلاف تصویر چندگانه را به عنوان ورودی می گیرد و از سه حالت انعطاف پذیر تشکیل شده است.

ماژول ها (آشکارساز مبتنی بر لنگر، و استراتژی مرتبط کردن جعبه های یک شی). آشکارساز اختلاف عمق را تخمین می زند

اشیاء را با تفاوت دید چپ و راست آنها شناسایی کرد

رویکرد دو مرحله‌ای می‌تواند ازنظر محاسباتی برای برنامه‌های کاربردی واقعی، که ظرفیت ذخیره سازی و محاسباتی محدودی دارند، گران باشد. رویکرد یک مرحله ای به طور مستقیم

احتمالات رده را پیش بینی می کند و جابجایی جعبه را با یک شبکه عصبی پیچشی تک رو به جلو میبرد. از این رو رویکرد یک مرحله ای تطابق بهتری بین سرعت و دقت دارد. آشکار سازهای شئ نماینده رویکرد یک مرحله ای هستند. یولو به طور مستقیم دسته بندی شی و جابجایی فضایی و مکان با یک شبکه پیچشی عصبی، با استنتاج سریع سرعت، را پیش بینی کرد. یولو از نرمال سازی دسته ای بعد از هر لایه پیچشی برای نتایج بهتر و از لایه های شبکه عصبی پیچشی به جای لایه های کاملا متصل برای طبقه بندی و رگرسیون افست مکان استفاده کرد. لیو و همکاران یک آشکارساز شی تک شات به نام اس اس دی را پیشنهاد کردند که اشیا را با استفاده از نقشه های ویژگی با گیرنده های مختلف پیش بینی میکند و زمینه عملیات را به اس اس دی برای زمینه اضافی و از مدل پیچیده تری برای دقت بهتر استفاده میکند. رتینا نت این مورد را بررسی کرد ،مشکل عدم تعادل طبقاتی شدید در یک مرحله فعلی رویکرد و آن را با طراحی مجدد تابع ضرر حل کرد.

ج. برآورد نابرابری

هدف از کار تخمین عمق پیش بینی نابرابری از هر پیکسل در تصویر ورودی است. چندین تخمین عمق روش ها با بهره مندی از سرعت توسعه شبکه های عصبی پیشرفت زیادی کرده اند. زبونتر و لکون شباهت های یک جفت تصویر استریو و شبکه عصبی پیچشی سیامی را محاسبه کردند. روش آنها الهام بخش چندین مطالعه درمورد تخمین عمق با استفاده از شبکه های عصبی پیچشی بود. دیسنپ تخمین عمق را به صورت نظارت شده به عنوان مشکل یادگیری و نابرابری های پیش بینی شده به طور مستقیم با شبکه پیچشی عصبی فرموله کرد. پی اس ام نت برای استفاده از ادغام ظرفیت اطلاعات جهانی و دستیابی به عملکرد پیشرفته از هرم فضایی استفاده کرد. روش های ذکر شده در بالا می توانند یک نقشه نابرابری دقیق ایجاد کنند اما آنها کند هستند و نیاز به محاسبات گسترده دارند. برای دستیابی به یک مبادله بهتر، انی نت عمق را در چند مرحله تخمین زد، که در طی آن مدل را میتوان در هر زمان، در بهترین تخمین فعلی آن، برای خروجی جستجو کرد. تک عمقی روشی بدون نظارت را پیشنهاد کرد که سعی در ایجاد یک نقشه نابرابری متراکم با آموزش شبکه با از دست دادن بازسازی تصویر داشت. برای آموزش فقط به جفت تصویر استریو نیاز داشت و شبکه را قادر ساخت تا اجرای تک تصویر و تخمین عمق با سرعت بیشتر را یاد بگیرد.

چارچوب آشکارساز نابرابری

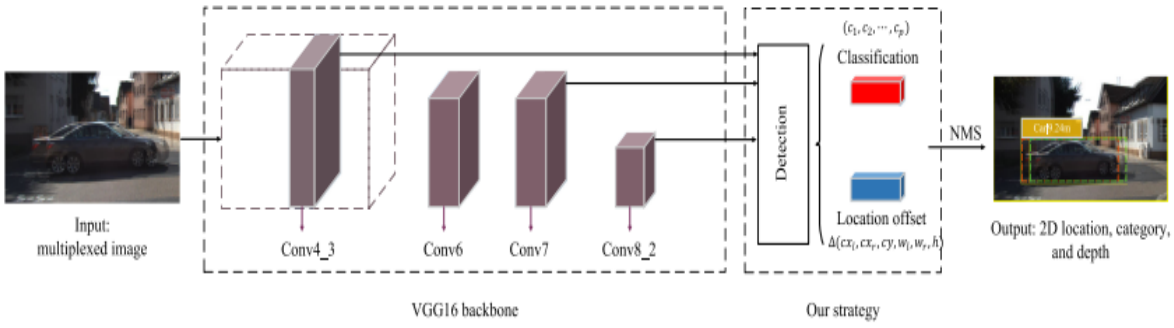
در این بخش ما یک استراتژی پیشنهاد میکنیم که میتواند با هر آشکارساز شی مبتنی بر لنگر برای تشکیل آشکارساز نابرابری ما همکاری کند. آشکارساز نابرابری میتواند به طور همزمان اشیا را تشخیص دهد و عمق اشیاء شناسایی شده در تصویر چندگانه را تخمین بزند. در ابتدا ما ویژگی های تصویر چندگانه را تجزیه و تحلیل میکنیم و دلیل این که چرا آشکارسازهای اخیر برای تصاویر چندگانه مناسب نیستند را توضیح میدهیم. سپس ما آشکارساز نابرابری را معرفی میکنیم که از شبکه استقامت، یک آشکارساز شئ مبتنی بر لنگر و استراتژی پیشنهادی ما تشکیل شده است.

ویژگی تصویر چندگانه

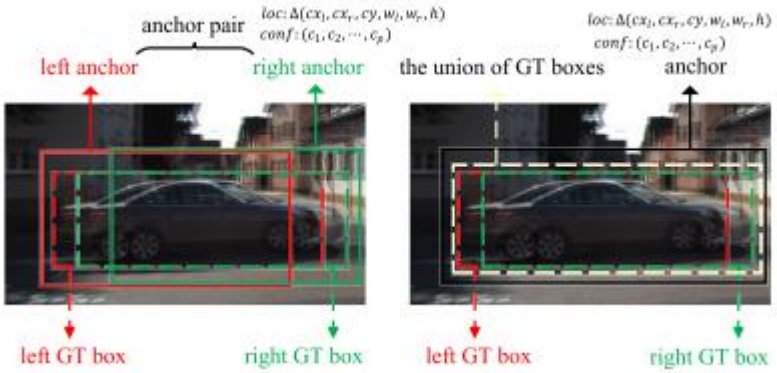
از آنجایی که تصویر چندگانه در این مقاله استفاده شده، ترکیبی از تصاویر از دو نمای افقی است. هر شئ در آ« دارای دو بخش است که میتوانند با یک جفت مرز افقی جعبه ها قرار گیرند. (جعبه های چین دار در شکل 5). نابرابری یک شئ را میتوان با فاصله پیکسل افقی بین مرکزهای دو جعبه آن تخمین زد. در دید استریو، فرمول اختلاف و عمق عبارت است از:

$$\text{عمق} = \frac{b \times f}{\text{نابرابری}}$$

که در آن (بی) فاصله خط پایه استریو و (ف) فاصله کانونی طول دوربین است. بهره مندی از این ویژگی، تصویر چندگانه امکان آشکارسازی مشترک و تخمین عمق را فراهم میکند. با این حال آشکارسازهای شئ مبتنی بر سیستم عصبی پیچشی کنونی، زمانی که مستقیما برای تشخیص اشیا در تصاویر چندگانه استفاده میشوند، محدودیت دارند. اطلاعات جفت هر دو شئ جعبه ها را نمیتوان برای آموزش وارد شبکه کرد. چپ یا راست، جعبه همان شئ ممکن است در طول سرکوب غیر حداکثر به دلیل منطقه همپوشانی بزرگ آنها فیلتر شود. از جانب خروجی یک آشکار ساز، مجموعه ای از جعبه های مرزی پیش بینی شده، نابرابری هر شئ در دسترس نیست زیرا آشکارساز نمیتواند کادرهای چپ و راست یک شئ را به هم مرتبط کند.



شکل 4. معماری شبکه دقیق آشکارساز نابرابری پیشنهادی



شکل 5. استراتژی مرتبط کردن جعبه های یک شی. سمت چپ: جفت لنگر استراتژی پیشنهادی؛ راست: استراتژی پیشنهاد شده توسط19

آشکارساز نابرابری

با توجه به مبادله سرعت در مقابل دقت، نابرابری آشکارساز ما بر اساس (درایو حالت جامد) یک آشکارساز یک مرحله ای است. درایو حالت جامد بر روی یک شبکه اصلی ساخته شده است که با چند لایه شبکه عصبی پیچشی به پایان میرسد (یا به انتهای آن کوتاه شده است). برای شناسایی اجسام با اندازه های مختلف، درایو حالت جامد، از نقشه های ویژگی با زمینه های دریافتی مختلف، برای پیش بینی امتیازات و آفست ها برای لنگرهای از پیش تعریف شده، استفاده میکند.این پیش بینی ها توسط فیلترهای بعدی کانال سه در سه انجام میشود، یک فیلتر برای امتیاز طبقه بندی و یکی برای جابجایی مکان از لنگرها. در نهایت سرکوب غیر حداکثری، برای کاهش افزونگی و به دست آوردن نتایج تشخیص، استفاده میشود. جزییات بیشتر را میتوان در 5 یافت. برای فعال کردن آشکارساز، برای شناسایی و مرتبط کردن سمت چپ و کادرهای سمت راست از همان شئ، در تصویر چندگانه ما یک استراتژی به نام جفت لنگر پیشنهاد میکنیم و با آن همکاری میکنیم که با استقامت و آشکارساز درایو حالت جامد شکل میگیرد. آشکارساز نابرابری ما معماری شبکه دقیق است که در شکل 4 نشان داده شده است.

1) جفت لنگر

با الهام از هر جسم دارای یک جفت افقی سمت چپ و جعبه های سمت راست (جی تی) جفت لنگر را پیشنهاد میکنیم که پسوند لنگر است. هر جفت لنگر از یک جفت تشکیل شده است، لنگرهای افقی چپ و راست همانطور کهه در شکل 5 نشان داده شده است. برای هر جفت لنگر (ال او یو) لنگر چپ آن، (ال او یو ال) را محاسبه میکنیم. با جعبه (جی تی) سمت چپ و لنگر راست آن (ال او یو آر) با جعبه (جی تی) سمت راست مربوطه. اگر (ال او یو ال) (ال او یو آر) آن باشد هر دو بالای 0.5 یک برچسب مثبت به لنگر اختصاص داده میشوود. اگر (ال او یو ال) و (ال او یو آر) هر دو جفت باشند یک برچسب منفی زیر 0.5 اختصاص داده میشود. هر جفت لنگر یک امتیاز طبقه بندی را پیش بینی میکند به

طوری که لنگرهای چپ و راست آن درجه طبقه بندی را به اشتراک میگذارند. ما به جفت لنگر مثبت اجازه میدهیم جابجایی مکان را پیش بینی کند. نسبت به سمت چپ و جعبه های (جی تی) سمت راست جایی که از (سی وای) (سی ایکس) برای نشان دادن افقی استفاده میکنیم و مختصات عمودی مرکز کادر در فضای تصویر برای عرض و ارتفاع جعبه، برای عبارت مربوطه در کادر چپ و راست. توجه داشته باشید که از همان (سی وای) استفاده میکنیم که (اچ سی وای 1) را جبران میکند. برای سمت چپ (اچ 1) و جعبه های مناسب استفاده میکنیم زیرا ما از تصاویر استریو اصلاح شده برای شبیه سازی تصاویر چندگانه استفاده میکنیم. بنابراین برای هر شش افست جفت لنگر به جای چهار در درایو حالت جامد اصلی را داریم.

از آنجایی که هر کدام از جعبه های چپ و راست شی پیش بینی شده توسط همان جفت لنگر و به اشتراک گذاشته شده امتیاز طبقه بندی آنها هستند، به طور طبیعی به عنوان یک دسته مرتبط است.

ما از (ان ام اس) برای پیش بینی کردن استفاده میکنیم. جعبه چپ و راست اشیا به طور جداگانه برای کاهش افزونگی و دریافت و تشخیص نتایج نهایی هستند. یک شی پیش بینی شده نگهداری خواهد شد اگر کادر چپ و راست هر دو بعد از (ان ام اس) نگه داشته شوند.

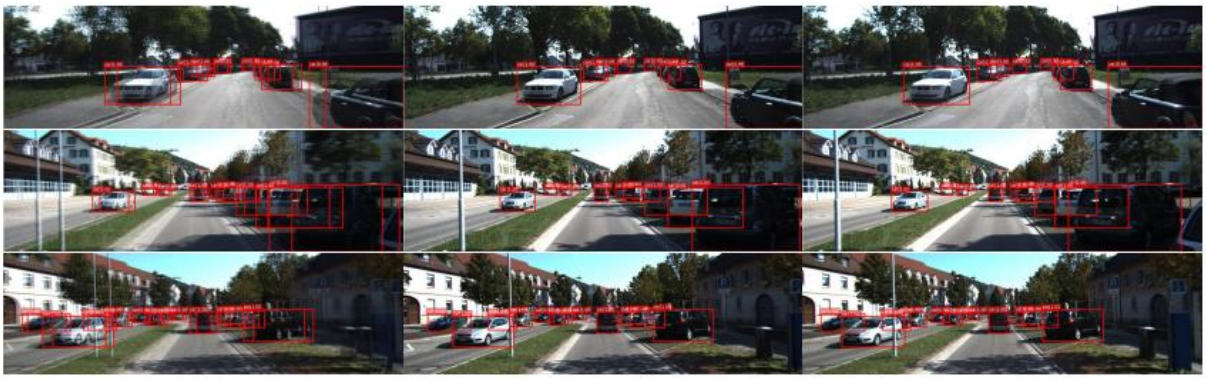
2) تفاوت با استراتژی فعلی

استریو (آر سی ان ان) یک استراتژی ساده اما خشن را پیشنهاد کرد. به عنوان استریو استراتژی، به جعبه های مشابه هدف – شئ، اشاره میشود.

همان طور که در شکل 5 نشان داده شده است، استریو استراتژی اتحادیه را به جعبه های حقیقت زمینی چپ و راست که به آن اتحادیه (جی تی) گفته میشود، و جعبه به عنوان هدف برای طبقه بندی اشیاء، اختصاص داده است. و یک برچسب مثبت به یک لنگر اختصاص داده میشود اگر (ال او یو) با یکی از جعبه های (جی تی) اتحادیه بالاتر از آستانه (تی اچ) باشد یا یک برچسب منفی اگر (ال او یو) زیر (تی ال) باشد. هر لنگر مثبت با توجه به جبران خسارات به سمت چپ و راست جعبه های (جی تی) موجود در اتحادیه هدف جعبه (جب تی) پیش بینی میکند. با این حال لنگر مثبت دارای یک (ال او یو) در بالاست. (تی اچ) با جعبه اتحاد نمیتواند تضمین کند که آن نیز دارای یک (ال او یو) بالا برای هر جعبه داخل جعبه اتحادیه است. به عبارت دیگر لنگر با برچسب مثبت ممکن است (ال او یو) زیر (تی اچ) داشته باشد یا حتی زیر (تی ال) با کادر چپ یا راست باشد.

Method	Setting	AP_{left}			AP_{stereo}			Runtime(s)
		Easy	Mode	Hard	Easy	Mode	Hard	
Faster R-CNN [30]	-	99.23	98.40	90.88	-	-	-	0.082
MFFD [44]		91.16	84.01	72.43	-	-	-	0.005
YOLOv3 [37]		95.96	95.51	88.26	-	-	-	0.031
SSD [5]		98.78	96.06	88.49	-	-	-	0.027
Disparity Detector	*strategy stereo* [19]	98.37	95.54	85.93	93.17	90.37	81.15	0.027
	anchor pair (ours)	98.55	96.00	88.62	93.69	91.87	83.14	0.027

جدول 1. دقت متوسط (در درصد) تشخیص، ارزیابی شده در مجموعه ارزیابی KITTI



شکل 6. نمونه هایی از نتایج تشخیص در مجموعه
با استفاده از روش پیشنهادی. سمت چپ: نتایج تشخیص در تصاویر چندگانه. وسط: نگاشت نتایج KITTI ارزیابی
ل ; راست: نگاشت نتایج تشخیص به تصاویر سمت راست | تشخیص به تصاویر سمت چپ

جزئیات پیاده سازی

شکل لنگر

Cluster IoU متفاوت از پیاده سازی در شکل 5، شکل لنگر توسط الگوریتم (کا مینز) پیشنهادی تعیین میشود. ما ابتدا از اشکال مختلف (در این مقاله 6 = n1) n1 را بر روی اجرا می کنیم مجموعه آموزشی برای انتخاب خودکار [36] k-mean یک الگوریتم Cluster IoU .توسط استریو استراتژی استفاده می شود n1 و این اشکال لنگر .(w, h) لنگر (لنگر، مرکز) IoU - (لنگر، مرکز) = 1 d :است با متریک فاصله

سپس در هر یک از خوشه های

k-means از n1 استفاده میکنیم.

با فاصله اقلیدسی برای به دست آوردن

n2 (n2 = 4

در این مقاله از فواصل مختلف

n1 × n2

جفت لنگر برای جفت لنگر استراتژی پیشنهادی

شبکه

ما هنگام اجرای مجدد تغییرات جزئی ایجاد کرده ایم. (اس اس دی) اندازه ورودی شبکه 576 ضرب در 320 است لایه های بعد از (کنو 8) را در (اس اس دی) اصلی حذف میکنیم و از نقشه ویژگی سه لایه (کنو 3 و 8 و 2 و 7) برای پیش بینی استفاده میکنیم. تنظینات دیگر مانند افزایش داده و استخراج نمونه سخت به عنوان (اس اس دی) اصلی یکسان هستند. ما شبکه را با استفاده از (اس جی دی) با یک کاهش وزن آموزش میدهیم. ما 100هزار تکرار (دسته اندازه در مجموع 16) در یک (تی آی) است. میزان یادگیری در ابتدا روی 0.001 تنظیم شده و با ضریب 0.1 کاهش میابد. 60 و 80 کا تکرار.

آزمایشات

در این بخش آشکارساز نابرابری پیشنهادی را در مجموعه داده تشخیص، ارزیابی میکنیم. در ابتدا آماده سازی مجموعه داده را معرفی میکنیم سپس استراتژی پیشنهادی خود را با استراتژی استریو (آر سی ان ان) در مورد عملکرد تشخیص شی و تخمین عمق به ترتیب مقایسه میکنیم.

آماده سازی مجموعه داده

مجموعه داده تشخیص (کی آی تی تی آی) آموزش 7481 را ارائه میدهد. جفت های تصویر استریو و برچسب جعبه محدود کننده سه بعدی (جعبه 2 بعدی برچسب تصویر چپ یا راست را میتوان با نمایش جعبه 3 بعدی به تصویر مربوطه محاسبه کرد. ما تصاویر چندگانه را با استفاده از جفت تصویر استریو شبیه سازی میکنیم به دنبال استریو (آر سی ان ان) این مقاله فقط از برچسب های دسته خودرو برای آموزش و ارزیابی استفاده میکند از بقیه تصاویر برای مجموعه ارزیابی استفاده میشود. ارزیابی دارای سه سطح دشواری است... آسان متوسط و سخت که بر حسب انسداد اندازه و سطوح برش اجسام تعریف میشوند. بررسی 20 برای یک تعریف دقیق از سطوح دشواری. ما همچنین از (ک یآی تی تی آی) استریو برای آموزش عمق استفاده میکنیم. روش های برآورد 39 و 41 و 43. این شامل 200 جفت تصویر استریو آموزشی با اختلافات نادر از واقعیت است که با استفاده از (لیدار) به دست آمده است.

عملکرد تشخیص شی

در این بخش آشکارساز های نابرابری پیشنهادی عملکرد تشخیص اشیا را ارزیابی میکنیم. ما نابرابری را با دو استراتژی در مجموعه آموزش تصویر چندگانه و ارزیابی آنها بر روی مجموعه ارزیابی چندگانه را آموزش میدهیم. ما همچنین آشکارساز پایه (اس اس دی) و غیره را آموزش میدهیم. آشکارساز های رایج فستر آر – ام اف اف دی 44 – سی ان ان 30 – یولو 37 – در سمت چپ مجموعه آموزش تصویر و ارزیابی آن در سمت چپ مجموعه ارزیابی تصویر است. برای فستر آر سی ان ان، اندازه تصویر اصلی به 600 پیکسل در سمت کوتاه تر تغییر میکند. برای اس اس دی – ام اف اف دی – یولو 3 اندازه تصویر ورودی به 576 ضرب در 320 تغییر داده شده است. همه آشکارساز ها همان شکل لنگر معرفی شده را دارند. هدف آشکارساز نابرابری ما، شناسایی همزمان و مرتبط کردن جعبه های مربوط به همان شی در تصویر چندگانه است. علاوه بر ارزیابی میانگین دقت (ای پی) در تصویر سمت چپنگاشت نتایج شخیص به تصویر سمت چپ و ما همچنین از متریک (ای پی) استریو استفاده میکنیم که در استریو آر سی ان ان 19 برای ارزیابی عملکرد ارتباط استفاده میشود. در استریو (ای پی) یک جفت جعبه چپ و راست به عنوان (ترو) در نظر گرفته میشود. در صورت رعایت شرایط زیر، (تی پی) مثبت است.

1) حداکثر (ال او یو) بین کادر سمت چپ و (جی تی) چپ جعبه بالاتر از آستانه است.

2) حداثر (ال او یو) بین جعبه سمت راست و (جی تی) سمت راست، جعبه بالاتر از آستانه است.

3) جعبه های انتخاب شده چپ و راست (جی تی) متعلق به هدف و شئ یکسان هستند.

ما بهترین روش را با رنگ قرمز پررنگ مشخص میکنیم. همانطور که گزارش شده است، در جدول 1 جفت لنگر پیشنهادی با حاشیه های زیاد عملکرد بهتری از استراتژس استریو 19 دارد. به طور خا ص پیشنهاد شده است، جفت لنگر بیش از 2.69 از استراتژی استریو بهتر عمل میکند و 1.99 درصد برای سطح سخت. ما به ترتیب آ« را به تطابق دقیق استراتژی خود با جعبه های لنگر و (جی تی) نسبت میدهیم. برخی از نمونه های تشخیص نابرابری آشکارساز در شکل 6 نشان داده شده است. ما همچنین مشاهده میکنیم که آشکارساز نابرابری ما، که ورودی آن تصویر چندگانه است، میتواند در مقایسه با آشکارساز های معمولی که تصویر سمت چپ را بع عنوان وردی میگیرند، عملکرد قابل مقایسه ای داشته باشد. وظیفه ما(تشخیص اشیا از دو دیدگاه) چالش برانگیز تر و دشوار تر از تشخیص اشیا روی تصویر سمت چپ با استفاده از آشکارساز مشترک است. در اطلاعات بصری یک شئ در نمای چپ و راست نامتعادل میباشد. به طور خاص برخی از اشیا قابل مشاهده در نمای چپ را میتوان به طور کامل مسدود کرد (حتی نامرئی) در نمای ست راست همانطور که در شکل 7 نشان داده شده است. به دلیل کمبود بصری اطلاعات در نمای درست، روش ما نمیتواند این اشیا را شناسایی کند، این مشکل عدم تعادل اطلاعات بصری تمرکز اصلی کار آینده ما خواهد بود. در این بخش، تفاوت پیشنهاد عملکرد آشکارساز برای تخمین عمق سطح شئ را ارزیابی میکنیم. ما نتایج آشکارساز نابرابری را با دو

استراتژی در مجموعه ارزیابی گزارش میکنیم. برای ارزیابی از نقطه پایانی (ای پی ای) استفاده میکنیم که به عنوان میانگین فاصله اقلیدسی بین نابرابری برآورد شده و حقیقت زمین محاسبه میشود. ما همچنین از درصد نابرابری با (ای پی ای) بزرگتراز پیکسل (تی) استفاده میکنیم. در اینجا یک شئ در صورتی صحیح در نظر گرفته میشود که اختلاف (ای پی ای) کمتر از پیکسل (تی) باشد. و نابرابری یک شئ با فاصله پیکسل افقی بین مرکز آن در کادر سمت چپ و کادر سمت راست تخمین زده میشود.

جدول 2 نتایج مقایسه اجسام با سطوح انسداد متفاوت را نشان میدهد. ما بهترین روش را با رنگ قرمز پررنگ مشخص میکنیم.

در برچسب (کی آی تی تی آی) اشیا با انسداد 0 به طور کامل قابل مشاهده هستند. اکولوژن=1 به این معنی است که اشیا تا حدی مسدود شده اند. و اشیا با انسداد=2 تا حد زیادی مسدود هستند. جفت لنگر ما بهتر از استراتژی استریو در همه شئ است. سطوح انسداد به جز>5 پیکسل است زمانی که انسداد = 2 باشد.

این نشان میدهد که آشکارساز با استراتژش پیشنهادی ما توانایی بهتری برای مکان یابی اشیا دارد. ما کمی ننتایج تخمین عمق آشکارساز نابرابری با جفت لنگر در شکل 9 نشان میدهیم.

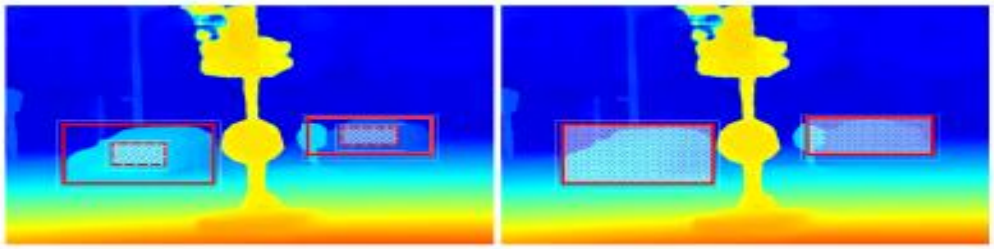
شکل 7. اطلاعات بصری نامتعادل شی در نماهای مختلف. سمت چپ: اشیاء در تصویر چندگانه. وسط: اشیاء در تصویر سمت چپ.

اشیاء در تصویر سمت راست :

TABLE 2. The proposed Disparity Detector's performance of object-level depth estimation.

Method	Setting	Occlusion=0				Occlusion=1				Occlusion=2			
		EPE	>1px	>3px	>5px	EPE	>1px	>3px	>5px	EPE	>1px	>3px	>5px
Disparity Detector	*strategy stereo* [19]	1.23	41.78	7.46	2.71	1.37	48.80	9.12	2.76	1.46	54.38	10.59	3.14
	anchor pair (ours)	1.11	35.44	6.58	2.16	1.26	42.77	8.00	2.67	1.38	48.15	9.54	3.28

.جدول 2. عملکرد آشکارساز اختلاف پیشنهادی در برآورد عمق سطح شی



شکل 8. ترکیب روش های تشخیص شی و تخمین عمق به پیش بینی عمق سطح شی جعبه مرزی (قرمز) جسم است پیش‌بینی می‌شود و روش‌های تخمین عمق خروجی می‌دهند نقشه نابرابری متراکم سمت چپ: با استفاده SSD [5] توسط از اختلاف میانگین پیکسل های مرکزی به عنوان نابرابری شی راست: استفاده از اختلاف میانه پیکسل ها در داخل جعبه به عنوان نابرابری شی.

با استفاده از آشکارساز اختلاف (با KITTI شکل 9. نمونه هایی از نتایج تخمین عمق در سطح شی در مجموعه ارزیابی جفت لنگر). عدد به رنگ قرمز پررنگ

.عمق حقیقت زمین است و عدد به رنگ زرد پررنگ عمق پیش‌بینی‌شده از روش ما است

Method	Setting	Occlusion=0			Occlusion=1			Occlusion=2			Runtime(s)
		EPE	>3px	>5px	EPE	>3px	>5px	EPE	>3px	>5px	
Disparity Detector	*anchor pair*	1.11	6.58	2.16	1.26	8.00	2.67	1.38	9.54	3.28	0.027
SSD [5]+MC-CNN [39]	*median*	2.11	22.11	11.72	3.44	34.83	19.42	7.62	77.54	55.46	0.704
	mean	1.62	14.75	7.91	3.55	33.88	20.67	7.78	80.44	57.08	
SSD [5]+Monodepth [43]	*median*	1.69	16.44	5.45	2.87	32.71	14.80	6.21	69.46	44.91	0.072
	mean	1.70	17.34	5.61	3.33	36.67	18.55	6.80	74.11	49.94	
SSD [5]+PSMNet [41]	*median*	0.98	7.53	3.27	1.72	15.12	7.79	5.98	62.74	43.74	0.612
	mean	0.93	7.10	3.39	2.34	23.03	12.61	6.61	73.19	47.07	

جدول 3. عملکرد تخمین عمق در سطح شی با ترکیب تشخیص شی و تخمین عمق

ما همچنین یک آزمایش جالب انجام میدهیم که با روش های تخمین عمق 39 – 41 و 43 برای پیش بینی عمق سطح شئ ترکیب میشود. این روش های تخمین عمق هستند که با مجموعه داده استریو (کی آی تی تی آی) آموزش دیده اند. ما از جعبه های محدود کننده (اس اس دی) برای مکان یابی اشیا و دریافت نابرابری هر شئ از نقشه نابرابری پیش بینی شده بر اساس عمق روش های برآورد استفاده میکنیم. همانطور که در شکل 8 نشان داده شده است، نابرابری یک شئ را میتوان با میانگین اختلاف مرکزی آن پیکسل ها یا اختلاف میانه پیکسل ها در محدوده آن جعبه تخمین زد.

$$disp_{mean}^{object} = \frac{1}{0.4w \times 0.4h} \times \sum_{cx-0.2w}^{cx+0.2w} \sum_{cy-0.2h}^{cy+0.2h} disp(i,j) \quad (3)$$

$$disp_{median}^{object} = median\{disp(i,j)|(i,j) \in BB\} \quad (4)$$

که در آن به ترتیب (سی وای) ، (سی ایکس) ، (اچ)، (دبلیو) ، عرض، ارتفاع و مرکز جعبه مرزی شئ (بی بی) هستند. دیسپ نقشه نابرابری با روش تخمین عمق پیش بینی شده است به عنوان ام سی پی اس و مونودپ.

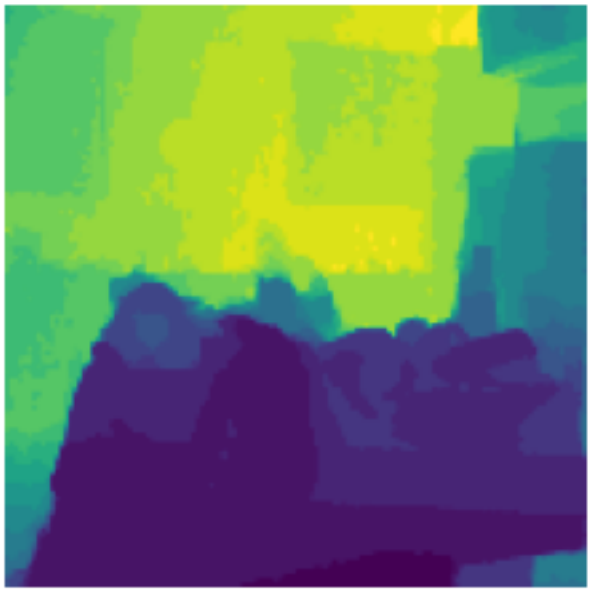
جدول 3 نتایج ارزیابی را در مجموعه ارزیابی تشخیص (کی آی تی تی آی) نشان میدهد. ما بهترین روش را با خط برجسته مشخص میکنیم. میتوان مشاهده کرد که این طرح ترکیبی، با پیشرفته ترین روش تخمین عمق، هنگامی که اشیا مسدود نشده اند انسداد = 0 ، عملکرد قابل مقایسه ای را به دست می آورد.

با این حال، هنگامی که جسم مسدود میشود، عملکرد آنها به شدت افت میکند انسداد = 1 و انسداد= 2

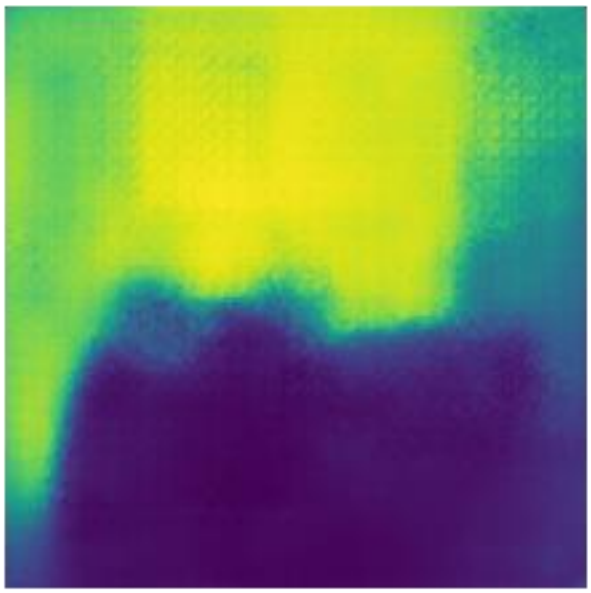
زیرا اکثر پیکسل ها در محدوده قرار دارند و جعبه به شئ تعلق ندارد. در مقابل پیشنهاد ما آشکارساز نابرابری در تمام انسداد ها به خوبی و به طور پیوسته عمل میکند. علاوه بر این در روش ما، سطوح زمان بسیار کمتری مصرف میکنند. 3 ضرب در 20 ضرب در سرعت.

شبکه‌ی عصبی Estimator:

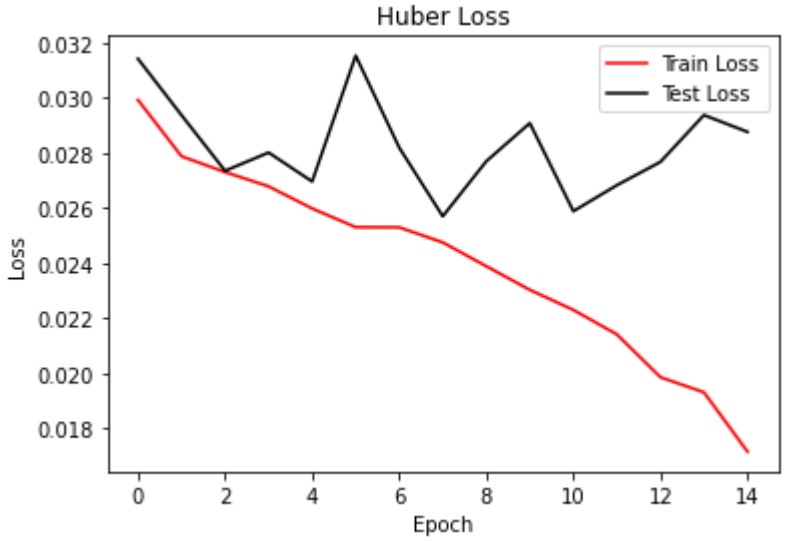
نحوه عملکرد شبکه‌های این پروژه به این صورت بوده که به طول کلی دو شبکه عصبی وجود دارند، شبکه estimator وظیفه‌ی تخمین عمق اشیا در تصویر را به عهده دارد. این شبکه داده‌های تصویر خام را که به طور عادی روی دیسک قرار دارند خوانده و پس از پیش‌پردازش‌های لازم آنها را به داده‌های ماتریس تبدیل کرده و به عوان ورودی دریافت میکند. ورودی تصاویر ۱۲۸ در ۱۲۸ پیکسل با ۳کانال رنگی ((RGBهستند. بعد از پردازش‌های داخلی شبکه، خروجی این شبکه یک تصویر ۱۲۸ پیکسل در ۱۲۸ با یک بعد داده است که برای هر پیکسل در تصویر سعی شده میزان فاصله آن از دوربین تشخیص داده شود. در پایین نمونه‌ای از ورودی و خروجی مدل رابه همراه داده‌ی عمق واقعی مشاهده میکنید:

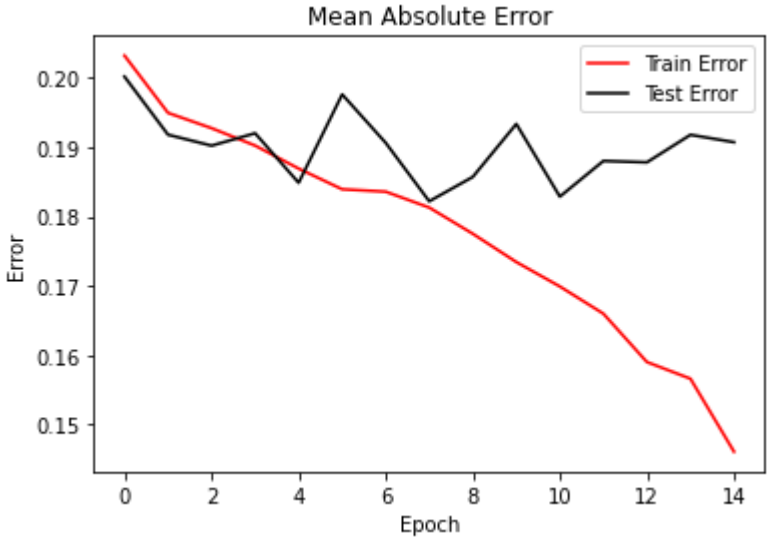


میزان پیش بینی شده توسط شبکه:

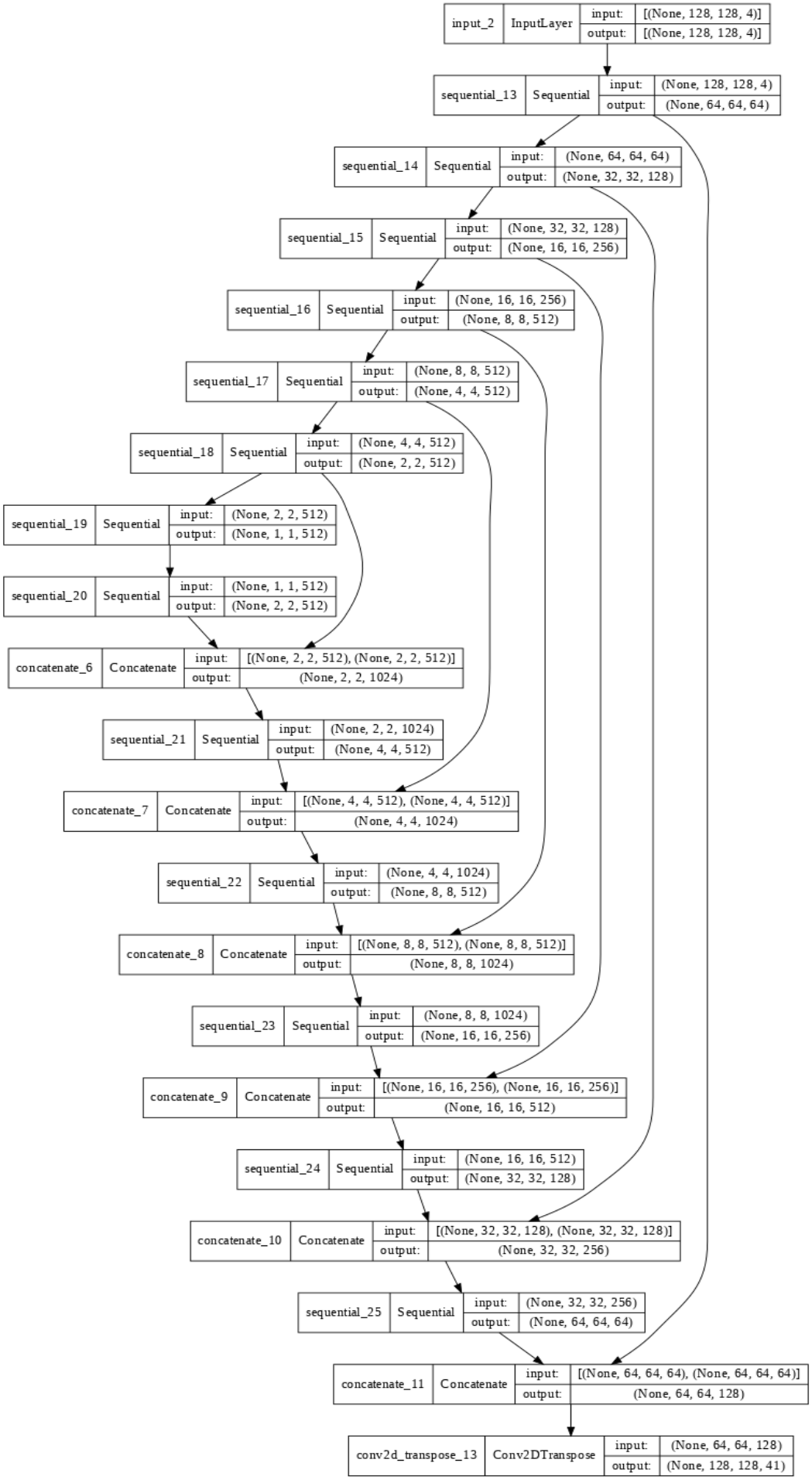


این شبکه‌ی عصبی با تابع بهینه‌ساز Adam ساخته شد. تابع خطایی که برای این شبکه‌ی عصبی میتوان استفاده کرد بیشمار هستند اما در این پروژه از تابع خطای Huber استفاده شده که خطای بین ماتریس هدف و ماتریس پیش‌بینی شده را به طرز ساده اما قدرتمندی پیش‌بینی میکند. برای متریک آموزش مدل از تابع MAE استفاده شده. همچنین برای آموزش مدل از روند Early Stopping استفاده شد با حد تحمل ۷ epoch، که در صورت overfit کردن روند آموزش مدل را به طور خودکار قطع کرده و به نقطه‌ی قبل از overfit باز میگردد. نتیجه نهایی مدل بعد از ۱۵ epoch آموزش برابر با خطای ۰،۰۱۷ برای تابع Huber و خطای ۰،۱۴ برای تابع MAE بود. همچنین عملکرد مدل روی دیتاست تست برابر با خطای 0.028 برای تابع Huber و مقدار خطای ۰،۱۹ برای تابع MAE بود که نتیجه بسیار مطلوبی است. نمودار آموزش مدل را میتوانید در زیر مشاهده کنید:

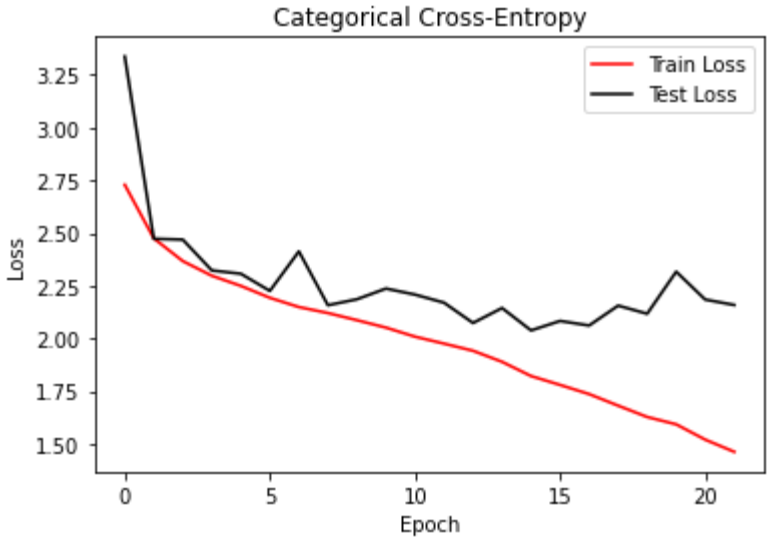


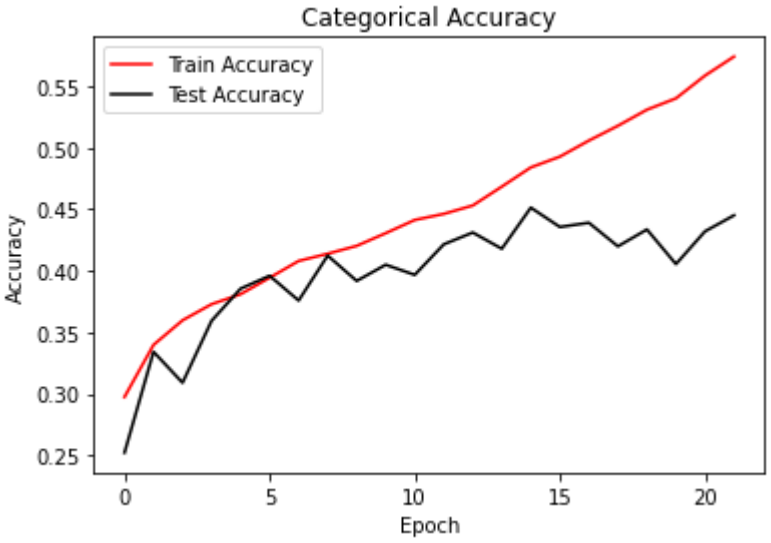


شبکه‌ی عصبی Detector:
در مرحله‌ی بعد شبکه‌ی عصبی detector قرار دارد. داده‌ی ورودی این شبکه‌ی عصبی همان تصاویر ورودی عادی که به شبکه‌ی عصبی قبلی داده شده‌اند هستند اما به آنها داده‌های عمق به عنوان بعد چهارم اضافه شده که شبکه‌ی عصبی علاوه بر داده‌های تصویر و رنگ، داده‌های عمق را نیز به عنوان ورودی دریافت کرده و خروجی یک ماتریس ۱۲۸ در ۱۲۸ پیکسل است که هر پیکسل دارای یکی از ۴۲ کلاسی هستند که در کلاس‌بندی‌های دیتاست NYU Depth V2 قرار دارند. پیکسلی که مقدارش برابر با 0 باشد کلاسی ندارد.
ساختار شبکه‌ای که در این مدل استفاده شده یک حالت تغییر یافته از شبکه ی عصبی Unet است. این شبکه‌ی عصبی درواقع یک generator یا GAN است که بوسیله لایه‌های کانولوشن دو بعدی تصاویر را دریافت کرده و تصاویر تغییر یافته را به عنوان خروجی میدهد. یک شبکه‌ی U-net از دو بخش تشکیل شده:
یک بخش انکودر یا upsampler و بخش دیگر دیکودر یا .downsampler
هر بلاک در انکودر از لایه‌های کانولوشن، batch normalizer و در نهایت تابع فعالساز ReLU تشکیل میشود. میزان فیلتر و تعداد فیلتر های استفاده شده در هر بلاک بسته به اینکه در کدام بخش از شبکه قرار دارد متفاوت است.
هر باک در بخش دیکودر از لایه‌های کانولوشن ترانهاده شده، batch normalizer، لایه‌های drop out و تابع فعال‌ساز leay relu تشکیل میشود.
همچنین بین هر انکودر و دیکودر اتصال‌های skip برقرار شده تا شبکه‌ی عصبی بتواند با یادگرفتن روابط جدید بین لایه‌های مختلف از overfit کردن جلوگیری کند.
به طور کلی از ۷ بلاک انکودر و ۶ بلاک دیکورد به علاوه یک لایه‌ی کانولوشن به عنوان لایه‌ی خروجی استفاده شده.
معماری کلی شبکه را میتوانید در گراف زیر مشاهده کنید:


input_2 | InputLayer | input: [(None, 128, 128, 4)] | output: [(None, 128, 128, 4)]
sequential_13 | Sequential | input: (None, 128, 128, 4) | output: (None, 64, 64, 64)
sequential_14 | Sequential | input: (None, 64, 64, 64) | output: (None, 32, 32, 128)
sequential_15 | Sequential | input: (None, 32, 32, 128) | output: (None, 16, 16, 256)
sequential_16 | Sequential | input: (None, 16, 16, 256) | output: (None, 8, 8, 512)
sequential_17 | Sequential | input: (None, 8, 8, 512) | output: (None, 4, 4, 512)
sequential_18 | Sequential | input: (None, 4, 4, 512) | output: (None, 2, 2, 512)
sequential_19 | Sequential | input: (None, 2, 2, 512) | output: (None, 1, 1, 512)
sequential_20 | Sequential | input: (None, 1, 1, 512) | output: (None, 2, 2, 512)
concatenate_6 | Concatenate | input: [(None, 2, 2, 512), (None, 2, 2, 512)] | output: (None, 2, 2, 1024)
sequential_21 | Sequential | input: (None, 2, 2, 1024) | output: (None, 4, 4, 512)
concatenate_7 | Concatenate | input: [(None, 4, 4, 512), (None, 4, 4, 512)] | output: (None, 4, 4, 1024)
sequential_22 | Sequential | input: (None, 4, 4, 1024) | output: (None, 8, 8, 512)
concatenate_8 | Concatenate | input: [(None, 8, 8, 512), (None, 8, 8, 512)] | output: (None, 8, 8, 1024)
sequential_23 | Sequential | input: (None, 8, 8, 1024) | output: (None, 16, 16, 256)
concatenate_9 | Concatenate | input: [(None, 16, 16, 256), (None, 16, 16, 256)] | output: (None, 16, 16, 512)
sequential_24 | Sequential | input: (None, 16, 16, 512) | output: (None, 32, 32, 128)
concatenate_10 | Concatenate | input: [(None, 32, 32, 128), (None, 32, 32, 128)] | output: (None, 32, 32, 256)
sequential_25 | Sequential | input: (None, 32, 32, 256) | output: (None, 64, 64, 64)
concatenate_11 | Concatenate | input: [(None, 64, 64, 64), (None, 64, 64, 64)] | output: (None, 64, 64, 128)
conv2d_transpose_13 | Conv2DTranspose | input: (None, 64, 64, 128) | output: (None, 128, 128, 41)

این شبکه‌ی عصبی با تابع فعالساز Softmax و تابع بهینه‌ساز Adam ساخته شده و از تابع Categorical Cross Entropy به عنوان تابع خطا و برای آموزش از متریک Categorical Accuracy به عنوان تابع دقت استفاده شد. با استفاده از Early Stopping شبکه‌ی عصبی به مقدار ۲۲ epoch آموزش داده شد و میزان خطای CCE برابر با ۱,۴۶ و دقت برابر با ٪۵۷,۴ برای داده‌ی آموزشی و برای داده‌ی تست میزان خطای CCE برابر با ۲,۱۵ و دقت برابر با ٪۴۴,۵ بود. نمودارهای آموزش شبکه را میتوانید در این قسمت مشاهده کنید:



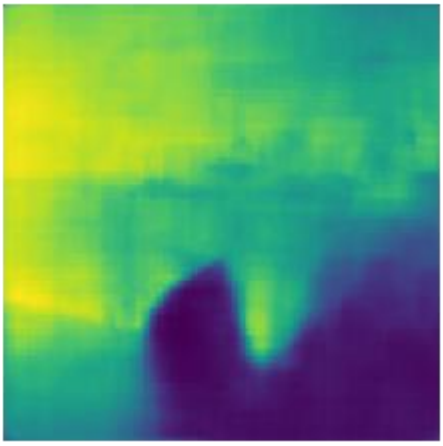


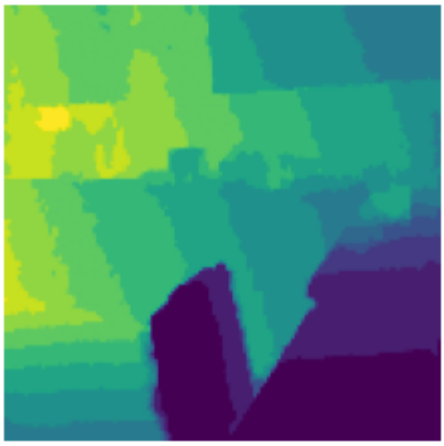
پیش‌بینی:

در نهایت چندین تابع نوشته شده که به طور جداگانه تصاویر را جهت پیش‌بینی از دیسک خوانده و بعد از پیش‌پرداز به شبکه‌ی estimator میفرستد. سپس از این مدل خروجی گرفته و عمق پیش‌بینی شده را با تصویر اولیه ادغام کرده تا یک تصویر ۴ بعدی مناسب شبکه‌ی detector بدست آورد. سپس این داده‌ی جدید به شبکه‌ی detector فرستاده شده و روی آن خروجی پردازش‌های لازم انجام میشود تا پیش‌بینی کلاس‌های مختلف شبکه‌ی عصبی روی قسمت‌های مختلف تصویر بدست بیاید. در این بخش یک نمونه از پیش‌بینی شبکه‌ی عصبی را مشاهده میکنید:

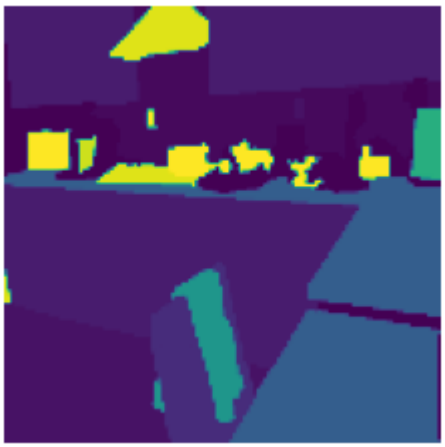
- عکس اولیه:

- داده‌ی عمق و پیش‌بینی شبکه عصبی:





- داده‌ی کلاس‌ها و پیش‌بینی شبکه عصبی:



- کلاس‌های پیش‌بینی شده توسط مدل:

wall
floor
cabinet
chair
bookshelf
counter
floor mat
refridgerator
lamp
bathtub
bag
other furniture