



دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک

عنوان فارسی

تشخیص قسمت های برجسته شی با مبدل تصمیم

استاد راهنما

دکتر پدram صالح پور

استاد مشاور

دکتر فرشی

پژوهشگر

آیناز رفیعی

زمستان 1403

از اساتید بزرگوارم جناب آقای دکتر پدرام صالح‌پور و جناب آقای دکتر فرشی برای تمام حمایت‌ها و زحمات بی دریغشان سپاسگزاری می‌کنم. از جناب آقای دکتر ----- که زحمت داوری این پایان‌نامه را به عهده داشتند سپاس فراوان دارم.

همچنین این پایان‌نامه را به پدر بزرگوار و مادر مهربانم تقدیم می‌کنم؛

بزرگترین و ارزشمندترین آموزگاران زندگی‌ام که همواره برایم تکیه‌گاه امن و مطمئنی بودند.

نام خانوادگی دانشجو: رفیعی نام: آیناز
عنوان پایان نامه: تشخیص قسمت های برجسته شی با مبدل تصمیم
استاد راهنما: دکتر پدram صالح پور استاد مشاور: دکتر فرش
مقطع تحصیلی: کارشناسی ارشد رشته: مهندسی کامپیوتر گرایش: هوش مصنوعی و رباتیکز دانشگاه: تبریز دانشکده: مهندسی برق و کامپیوتر تاریخ فارغ التحصیلی: 1403/10/10 تعداد صفحه: 89
واژگان کلیدی تشخیص قسمت های برجسته ی اشیاء، مبدل تصمیم، یادگیری تقویتی، Decision Transformer
<p>چکیده</p> <p>شناسایی اشیای برجسته در تصاویر یکی از موضوعات مهم در حوزه بینایی کامپیوتر است که کاربردهای گسترده‌ای در زمینه‌هایی مانند بازشناسی اشیاء، ردیابی اهداف و تحلیل تصاویر دارد. این پایان‌نامه به بررسی و توسعه یک مدل مبتنی بر Transformer به نام Decision Transformer برای انجام وظایف شناسایی اشیای برجسته در مجموعه داده DUTS پرداخته است. هدف اصلی این پژوهش، استفاده از قابلیت‌های ترنسفورمر در پردازش داده‌های پیچیده و ترکیب ویژگی‌های زمانی و تصویری برای افزایش دقت و کارایی در شناسایی اشیای برجسته است. مجموعه داده DUTS، که یکی از گسترده‌ترین مجموعه‌های داده در این حوزه محسوب می‌شود، شامل 10572 تصویر در بخش آموزش و 5019 تصویر در بخش آزمایش است. در این تحقیق، 8442 تصویر برای آموزش، 2111 تصویر برای اعتبارسنجی و 5019 تصویر برای آزمایش مورد استفاده قرار گرفته است. مدل Decision Transformer، که ابتدا برای مسائل یادگیری تقویتی طراحی شده بود، در این تحقیق برای شناسایی اشیای برجسته در تصاویر بازطراحی شده است. این مدل از ترکیب ویژگی‌های استخراج‌شده از تصاویر با استفاده از Vision Transformer (ViT) و اطلاعات زمانی بهره می‌برد. ساختار مدل شامل لایه‌های خودتوجه، شبکه‌های عصبی پیش‌خور و توابع فعال‌سازی است که امکان پردازش داده‌های چندوجهی و یادگیری توالی‌های پیچیده را فراهم می‌کند.</p> <p>برای مقایسه عملکرد مدل پیشنهادی، دو مدل دیگر نیز بررسی شده‌اند. مدل Visual Saliency Transformer یک روش مبتنی بر ترنسفورمر است که با طراحی معماری خودتوجه چندگانه برای تصاویر، برجستگی‌های بصری را با دقت بالا شناسایی می‌کند. این مدل با تمرکز بر بازنمایی دقیق ویژگی‌های محلی و جهانی تصویر، بهبود قابل توجهی در نتایج شناسایی اشیای برجسته ارائه می‌دهد. مدل دوم، Texture-guided Saliency Distilling for Unsupervised Salient Object Detection، از اطلاعات بافتی تصویر برای شناسایی اشیای برجسته به صورت بدون نظارت بهره می‌گیرد. این روش با استفاده از راهنمایی مبتنی بر بافت، برجستگی‌های تصویر را تقطیر کرده و نقشه‌های برجستگی دقیق‌تری تولید می‌کند.</p> <p>نتایج آزمایش‌ها نشان می‌دهد که مدل Decision Transformer در مقایسه با دو مدل دیگر در معیارهایی همچون MAE و F-measure عملکرد بهتری داشته است. این پژوهش توانایی‌های معماری ترنسفورمر را برای حل مسائل پیچیده بینایی کامپیوتر به نمایش گذاشته و نشان می‌دهد که Decision Transformer می‌تواند به عنوان ابزاری قدرتمند در شناسایی اشیای برجسته مورد استفاده قرار گیرد. با توجه به این نتایج، این تحقیق به عنوان یک گام مؤثر در گسترش استفاده از ترنسفورمرها در مسائل بینایی کامپیوتر شناخته می‌شود.</p>

فصل 1: کلیات تحقیق	1
1-1- مقدمه	2
2-1- بیان مساله	3
3-1- اهمیت و ضرورت تحقیق	4
4-1- اهداف تحقیق	6
5-1- اهداف کاربردی	7
6-1- سوالات تحقیق	8
7-1- فرضیات تحقیق	8
6-1- ساختار تحقیق	9
فصل 2: ادبیات نظری و پیشینه تحقیق	1
1-2- مقدمه	2
2-2- تشخیص قسمت های برجسته ی اشیاء	3
3-2- سیستم تشخیص قسمت های برجسته ی اشیاء	3
4-2- روش های تشخیص قسمت های برجسته ی اشیاء	6
1-4-2- پیش بینی تمرکز دید	7
2-4-2- تشخیص هم برجستگی	7
3-4-2- تشخیص عمق برجستگی تصاویر رنگی	7
4-4-2- پیش بینی نگاه اجتماعی	8
5-4-2- تشخیص شیء برجسته	8
6-4-2- مقایسه روش ها	8
5-2- مدل های سنتی تشخیص قسمت های برجسته اشیاء	10
1-5-2- روش های مبتنی بر کنتر است	10
2-5-2- روش های منطقه محور	10
3-5-2- روش های مبتنی بر پیشینه ی پیشین	11
4-5-2- روش های مبتنی بر انتقال	11
5-5-2- روش های مبتنی بر بهینه سازی	11
6-5-2- روش های هندسی و مبتنی بر فاصله	12
7-5-2- روش های ترکیبی	12
6-2- مدل های یادگیری عمیق تشخیص قسمت های برجسته اشیاء	13
1-6-2- شبکه های عصبی کانولوشنی	13
2-6-2- شبکه های عصبی بازگشتی	13
3-6-2- مدل های عمیق سلسله مراتبی	14
4-6-2- مکانسیم های مبتنی بر توجه	14
5-6-2- یادگیری تحت نظارت با ویژگی های عمیق	14
6-6-2- ویژگی های متنی و جهانی	15
7-6-2- ویژگی های چند مقیاسه	15

15	8-6-2- رویکرد های ترکیبی و بهینه سازی
15	7-2- مدل های مبدل تصمیم
18	8-2- اصطلاحات مدل های مبدل تصمیم
18	1-8-2- مسیر حرکت
18	2-8-2- بازگشت به رفتن:
19	3-8-2- مدل سازی دنباله ای
19	4-8-2- مکانیسم توجه
19	5-8-2- تعبیه حالت
20	6-8-2- اکشن جاسازی شده
20	7-8-2- سیاست شرطی پاداش
20	8-8-2- یادگیری تقویتی آفلاین
21	9-8-2- پنجره زمینه
21	10-8-2- مبدل علی
21	11-8-2- اکتشاف در مقابل بهره برداری
22	12-8-2- انتزاع زمانی
22	13-8-2- شبیه سازی سیاست
22	14-8-2- تنظیم دقیق
23	15-8-2- تعمیم
25	9-2- جمع بندی
27	فصل 3: روش شناسی تحقیق
28	1-3- مقدمه
28	2-3- معماری مدل پیشنهادی
32	1-2-3- مدل پیشنهادی
36	2-2-3- آموزش مدل
37	3-3- جمع بندی
40	فصل 4: یافته های تحقیق
41	1-4- مقدمه
41	2-4- دادگان
44	3-4- معیار های ارزیابی
44	1-3-4- معیار های مبتنی بر منطقه
45	2-3-4- معیار های مبتنی بر خطا
46	3-3-4- معیار های مبتنی بر ساختار
47	4-3-4- معیار های مبتنی بر مرز
47	5-3-4- ارزیابی کلی
48	4-4- ارزیابی مدل پیشنهادی
48	1-4-4- آزمایش اول
49	2-4-4- آزمایش دوم
51	3-4-4- آزمایش سوم (مدل مبدل تصمیم)
54	5-4- مقایسه روش مبدل تصمیم با روش های موجود
54	6-4- تحلیل نتایج

57	4-7- جمع‌بندی
59	فصل 5: نتیجه‌گیری و ارائه پیشنهاد
60	5-1- نتیجه‌گیری
60	5-2- کارهای آینده
63	مراجع
68	فهرست واژگان

صفحه

فهرست اشکال

4	شکل (2-1) تشخیص قسمت های برجسته ی شی
6	شکل (2-2) تفاوت تشخیص قسمت های برجسته ی شی روش های سنتی و مدرن
9	شکل (2-3) مقایسه ی رویکرد های مشابه تشخیص قسمت های برجسته اشیاء
24	شکل (2-4) رویکرد تشخیص قسمت های برجسته اشیاء با مبدل تصمیم
30	شکل (1-3) فرآیند کلی سیستم تشخیص قسمت های برجسته اشیاء
39	شکل (2-3) معماری مدل پیشنهادی
42	شکل (1-4) فراوانی داده ها
53	شکل (2-4) نتایج مدل پیشنهادی روی داده های آموزش و ارزیابی
57	شکل (3-4) مقایسه نتایج مدل پیشنهادی و مدل ViT و مدل A2S-v2 روی انواع داده

صفحه

فهرست جداول

38	جدول (1-3) مقادیر پارامترها
42	جدول (1-4) تقسیم بندی داده ها برای مجموعه داده ی DUTS
43	جدول (2-4) نمونه مجموعه داده ی DUTS

جدول (3-4) ماتریس درهمریختگی.....	45
جدول (4-4) نتایج مدل VST.....	49
جدول (5-4) نتایج مدل A2S-v2.....	50
جدول (6-4) نتایج مدل پیشنهادی برای مقادیر مختلف پارامترها.....	52
جدول (7-4) نتایج T-Test آزمایش سوم.....	53
جدول (8-4) نمونه مجموعه داده ی DUTS.....	56

فصل 1 : کلیات تحقیق

1-1- مقدمه

تشخیص قسمت های برجسته ی اشیاء¹ از جمله مسائل اساسی در حوزه ی هوش مصنوعی و تشخیص اجسام² است که در چند دهه اخیر توجهات گسترده ای را در قالب های عدیده به خود معطوف کرده است. تشخیص قسمت های برجسته ی اشیاء، قابلیت رایانه ای است برای درک چیزی که انسان با استفاده از سیستم بینایی برآورده میکند. شناخت رایانه از طریق شناخت اشیاء موجب می شود تا رایانه ها بتوانند ویژگی و نشانه های بصری مختلف، از جمله رنگ، کنتراست [1]، بافت و توزیع فضایی تحلیل کنند و اطلاعات مفیدی را از آن ها استخراج نمایند.

یکی دیگر از موارد مهمی که توسط شناخت اجسام رایانشی مورد مطالعه قرار گرفته ، کارایی محاسباتی در قسمت های برجسته ی اشیاء برای کاربرد های بی درنگ و مقیاس بزرگ است. هدف از تشخیص اشیاء برجسته این است که به طور خودکار این اشیاء مهم را در یک تصویر مشخص کند و سیستم های رایانه ای را قادر می سازد تا محتوای بصری را به طور مؤثرتری درک و تفسیر کنند، همچنین بهینه سازی پیچیدگی محاسباتی برای اجرای عملی مهم است. در بعضی مواقع در نظر گرفتن اطلاعات زمینه ای [3] پیرامون قسمت های برجسته ی اشیاء میتواند دقت تشخیص را بهبود ببخشد. نشانه های متنی ، مانند روابط فضایی و انسجام معنایی میتوانند به تمایز اشیاء برجسته واقعی از درهم رفتگی پس زمینه کمک کند [6,19]. این عملیات با یکپارچه شدن با سایر وظایف بینایی کامپیوتر، مانند تشخیص اشیاء، تقسیم بندی³ و طبقه بندی⁴، میتواند عملکرد سیستم را افزایش دهد.

در این تحقیق مبدل های تصمیم⁶ جهت فراهم کردن ویژگی های یادگیری تقویتی⁵ [21] برای ورودی های متوالی و کاربرد های بلادرنگ مطرح شدند و درحوزه ی تشخیص اشیاء برجسته مورد بررسی قرار خواهد گرفت. استفاده ی مستقیم از یادگیری تقویتی موجب انتشار خطا و برآورد بیش از حد میشود. درحالی که در مبدل های تصمیم، برخلاف یادگیری تقویتی، فرایند آموزش از روی یک

¹ Saliency Object Detection

² Object Detection

³ Segmentation

⁴ Classification

⁵ Reinforcement

⁶ Decision Transformer

⁷ Reward

⁸ Return

مجموعه پاداش های فوق العاده ⁷، بهترین را در نظر میگیرد به عبارت دیگر این روش میتواند بدون بازخورد ثابت ⁸ تنها با آموزش تجربیات قدیمی انجام شود. تعمیم این رویکرد جدید در تشخیص قسمت های برجسته ی اشیاء میتواند مورد توجه و بررسی قرار گیرد.[32]

1-2- بیان مساله

همانطور که در بخش 1-1 بیان شد، یکی از موارد مهمی که توسط شناخت اجسام رایانشی مورد مطالعه قرار گرفته، بحث استفاده از مبدل های تصمیم است که در راستای یادگیری تقویتی قرار میگیرد. یادگیری تقویتی توانایی تصمیم گیری و مدل کردن توالی طولانی و توزیع گسترده تر را دارد که موجب تعمیم و انتقال بهتر می شود. مدل کردن مبدل های تصمیم به فهم الگوی موجود در دیتا ها بدون نیاز به یادگیری سیاست ¹ یا بازخورد از روند تکامل ²، تنها با آموزش تجربیات قدیمی انجام می شود.[21] با این حال چگونگی بکار گیری روش مبدل های تصمیم در تقویت تشخیص قسمت های برجسته ی اشیاء یک مسئله ی چالش برانگیز است. تشخیص قسمت های برجسته ی اشیاء، بر خلاف نمونه سنتی خود، صرفاً به شناسایی اشیاء از پیش تعریف شده مربوط نمی شود، بلکه هدف آن برجسته کردن مناطقی است که توجه انسان را به خود جلب می کند. این عملیات پیش پردازش [27] نقشی اساسی در وظایف مختلف بینایی رایانه ای مانند تقسیم بندی تصویر، درک صحنه و بازیابی تصویر مبتنی بر محتوا دارد. با مشخص کردن مناطق با اهمیت بصری [32]، تشخیص قسمت های برجسته ی اشیاء، پردازش و تفسیر کارآمدتر تصاویر را تسهیل می کند و در نتیجه عملکرد کارهای پایین دستی را افزایش می دهد. در محیط های دنیای واقعی که قطعیتی وجود ندارد یا به معنای دیگر خروجی مدل ها همیشه قابل پیش بینی نیست، بکارگیری مبدل های تصمیم برای تشخیص قسمت های برجسته ی اشیاء یک مزیت پرکاربرد جهت تصمیم گیری، کنترل و بهینه سازی است.

تشخیص قسمت های برجسته ی شی در دستیابی به شناسایی دقیق و کارآمد در یک تصویر با چالش ها و مسائل خاص مرتبط اند. تصاویر اغلب حاوی صحنه های پیچیده با اشیا و مناطق متعدد هستند که میتوانند برای جلب توجه رقابت کنند. تمایز بین اشیاء برجسته واقعی و درهم و برهمی پس زمینه یا سایر عناصر بصری قابل توجه می تواند چالش برانگیز باشد [47]. از طرفی الگوریتم های تشخیص

¹ Policy

² Propagate return

اشیاء برجسته نیاز به پردازش مقادیر زیادی از داده های بصری و استخراج ویژگی های مرتبط دارند، که میتواند از نظر محاسباتی فشرده باشد. اطمینان از عملکرد و مقیاس پذیری برای مجموعه داده های بزرگ یک چالش مهم است. ادراک برجستگی میتواند ذهنی باشد و تحت تاثیر ترجیحات فردی، عوامل فرهنگی و زمینه باشد و توسعه الگوریتم هایی که روش های متنوعی را که در آن انسان ها برجستگی را درک میکنند، به تصویر می کشند [19]، یک کار پیچیده است. شناسایی قسمت های برجسته ی اشیاء اغلب مستلزم در نظر گرفتن اطلاعات زمینه ای، مانند روابط مکانی و انسجام معنایی است. ترکیب نشانه های زمینه ای درحالی که از اتکای بیش از حد ویژگی های زمینه ای خاص اجتناب می شود، یک مشکل غیر ضروری است. درک نیاز ها و ترجیحات کاربر، و همچنین توسعه رابط های بصری برای تعامل با محتوای بصری متنی بر برجستگی اشیاء، یک چالش طراحی در ایجاد راه حل های کاربر پسند و موثر است.

پرداختن به این مشکلات مستلزم تلاش های تحقیق و توسعه مداوم برای بهبود دقت، کارایی و قابلیت استفاده الگوریتم های تشخیص قسمت های برجسته اشیاء است که در نهایت عملکرد آنها را در برنامه های بینایی رایانه ای مختلف افزایش می دهد.

تشخیص قسمت های برجسته اشیاء را میتوان برای شناسایی و اولویت بندی مناطق مهم در تصاویر و ویدئو ها برای فشرده سازی، امکان ذخیره سازی و انتقال کارآمد محتوای بصری با حفظ کیفیت ادراکی استفاده کرد. ویژگی های مبتنی بر برجستگی میتوانند سیستم های بازیابی تصویر مبتنی بر محتوا را با امکان جست و جوی تصاویر براساس اشیاء یا مناطق برجسته بهبود بخشند و دقت و ارتباط را ارتقاء دهند.

با استفاده از فناوری تشخیص قسمت های برجسته اشیاء در این برنامه ها و دیگر برنامه ها، محققان و توسعه دهندگان میتوانند قابلیت های پردازش بصری را افزایش دهند، تجربیات کاربر را بهبود بخشند و راحل های نوآورانه را در دامنه های مختلف فعال کنند که بر مکانیزم های دقت و توجه بصری کارآمد متکی هستند.

1-3- اهمیت و ضرورت تحقیق

تشخیص اشیاء برجسته در طراحی رابط کاربر و تعامل انسان و رایانه نقش دارد، زیرا به ایجاد رابط هایی کمک می کند که عناصر بصری مهم را برای کاربران اولویت بندی و تأکید کنند. در دنیای

تکنولوژی برای هرچه آسان تر کردن یادگیری ماشین و منطبق کردن آن با خواسته های بدون درنگ انسان ضرورت برآوردن کردن نیاز های زیادی را پررنگ میکند. درک اشیاء برجسته در یک تصویر، تمایل طبیعی سیستم بینایی انسان به تمرکز بر عناصر مهم را تقلید می کند. با شناسایی قسمت های برجسته اشیاء، سیستم های کامپیوتری می توانند مکانیسم های توجه بصری را شبیه سازی کنند و پردازش اطلاعات مربوطه را در اولویت قرار دهند. شناسایی دقیق قسمت های برجسته ی اشیاء برای کارهایی مانند تقسیم بندی تصویر و طبقه بندی اشیاء ضروری است، جایی که تعیین و جداسازی اشیاء مهم از پس زمینه برای تجزیه و تحلیل و درک بیشتر ضروری است. در برنامه های ویژه مانند موتور های جستجوی تصویر و پایگاه داده های بصری اهمیت بیشتر استخراج قسمت های برجسته را نمایان میکند. از جمله مزایا و کاربرد در تسک های عمومی به صورت زیر است:

- امروزه با پیشرفت فناوری رانندگی خودمختار نیازمندی به رویکردهای بهتر تشخیص اشیاء جهت حفظ امنیت سرنشین و عابران پیاده و همچنین بهبود کیفیت و کارایی سیستم بیشتر حس می شود. تشخیص اشیاء برجسته می تواند به وسایل نقلیه خودران کمک کند تا اشیاء مهم در جاده مانند عابران پیاده، وسایل نقلیه و علائم راهنمایی و رانندگی را شناسایی و ردیابی کنند تا ایمنی و تصمیم گیری را در سناریوهای رانندگی در زمان واقعی بهبود بخشد. در مزیتی دیگر میتواند به تشخیص بهتر چاله، خرابی و دست انداز های آسفالت ها بپردازد تا راننده را از وجود هرگونه خطر احتمالی باخبر کند و سلامتی سرنشین هارا تضمین کند.
- در حوزه ی تصویر برداری پزشکی همچنان شاهد یکسری خطا های بصری از سوی پزشکان هستیم که میتواند در صورت عدم تشخیص یا خطا در تشخیص، بیماری مراجعه کننده را به سمت خطر بیشتری سوق دهد. تشخیص اشیاء برجسته می تواند به تجزیه و تحلیل تصاویر پزشکی، مانند اشعه ایکس، اسکن MRI، و اسلایدهای آسیب شناسی، با برجسته کردن مناطق مربوطه برای تشخیص، برنامه ریزی درمان، و اهداف تحقیقاتی کمک کند و احتمال خطا را در حد صفر برساند.
- الگوریتم های ردیابی اشیاء مبتنی بر برجسته بودن می توانند دقت و استحکام ردیابی اجسام متحرک را در ویدیوها برای نظارت، تجزیه و تحلیل ورزشی و سایر برنامه های نظارتی بهبود بخشند.

بنابراین استفاده از یک سیستم برای تشخیص قسمت های برجسته اشیاء، می تواند در جهت سرعت بخشیدن به کارها مفید واقع شود. در این طرح پیشنهادی، هدف ایجاد یک سیستم برای تشخیص قسمت های برجسته اشیاء از دید رایانه ای و درک تصویر است، که ماشین ها را قادر می سازد اطلاعات معنی داری را از داده های بصری استخراج کنند و طیف وسیعی از وظایف را با دقت و کارایی بیشتر انجام دهند.

در این مطالعه سعی خواهد شد از متدهای مبدل تصمیم¹ استفاده شود که مبتنی بر تسک های یادگیری تقویتی است. با در نظر گرفتن کاربرد چشم گیر قسمت های برجسته ی اجسام در زمینه و صنایع مختلف دستیابی به این رویکرد یک نقش مهم و ضروری است.

1-4- اهداف تحقیق

همان طور که در بخش 1-3 بیان شد، با توجه به پیشرفت فناوری و ماشینی شدن بیشتر کارها، کمبود سیستمی که بتواند فرآیند تشخیص قسمت های برجسته اشیاء را انجام دهد، به شدت احساس می شود. بنابراین هدف از این تحقیق طراحی سیستمی برای تشخیص قسمت های برجسته اشیاء می باشد که با استفاده از مبدل های تصمیم که هسته اصلی سیستم است، انجام می شود. ادغام مبدل تصمیم در وظایف تشخیص بخش های برجسته ، رویکرد جدیدی را نشان می دهد که از نقاط قوت معماری مبدل، در مدیریت داده های متوالی و گرفتن وابستگی دوربرد استفاده می شود.

مبدل های تصمیم میتوانند به طور موثر روابط بین اشیاء مختلف و زمینه های آنها را در یک تصویر مدل کنند. این به درک نه تنها خود اشیاء، بلکه نحوه تعامل آنها با یکدیگر در یک صحنه کمک میکند. با در نظر گرفتن فرآیند تصمیم تشخیص به عنوان دنباله ای از تصمیمات، مبدل های تصمیم میتوانند اولویت بندی ویژگی ها یا مناطق خاصی را در یک تصویر براساس مکانیسم های توجه آموخته شده بیاموزند. این اجازه می دهد تا نقشه های برجسته تری پویا تر و آگاه از زمینه ایجاد شود. مبدل های تصمیم به دلیل توانایی خود در تعمیم خوب در وظایف و حوزه های مختلف شناخته شده اند. استفاده از مبدل های تصمیم میتواند استحکام مدل های تشخیص برجسته را افزایش دهد واز آنها درسناپو های مختلف مانند تشخیص برجستگی در صحنه های پیچیده با چندین اشیاء همپوشانی شده موثر باشد.

¹ Decision Transformer

یکی از علل عدم استفاده از این سیستم ها، پیچیدگی محاسباتی و نیازهای منابع مرتبط با بسیاری از الگوریتم های مبتنی بر یادگیری عمیق است. این روش ها اغلب به قدرت پردازش و حافظه قابل توجهی نیاز دارند که میتواند مانعی برای استقرار در برنامه های کاربردی بلادرنگ یا در دستگاه هایی با منابع محدود مانند تلفن های همراه یا سیستم های تعبیه شده باشد. بنابراین هدف از این تحقیق یافتن راه حلی با عملکرد کافی و بدون سربالایی محاسباتی بالا هستند که قابل اجرا در دنیای واقعی باشد.

1-5- اهداف کاربردی

سیستم طراحی شده میتواند در وسایل نقلیه ی خودران، تصویربرداری پزشکی و ... جهت سهولت در کارها، صرفه جویی در وقت و افزایش ایمنی و سلامتی، مورد استفاده قرار گیرد. چند مورد از این موارد به شرح زیر می باشد:

- وسایل نقلیه خودران: کمک به خودروهای خودران برای شناسایی و تمرکز بر روی اشیاء مهم در محیط، مانند عابران پیاده، علائم راهنمایی و رانندگی و سایر وسایل نقلیه، تضمین ناوبری ایمن تر.
- بازیابی تصویر مبتنی بر محتوا: بهبود موتورهای جستجو با اجازه دادن به آنها برای بازیابی تصاویر بر اساس ویژگی های برجسته بصری، بهبود دقت نتایج جستجو.
- تصویربرداری پزشکی: کمک به رادیولوژیست ها با برجسته کردن مناطق مورد علاقه در اسکن های پزشکی (به عنوان مثال، تومورها)، تسهیل تشخیص سریع تر و دقیق تر.
- واقعیت افزوده¹ و واقعیت مجازی²: سیستم های واقعیت افزوده و مجازی را قادر می سازد تا بر روی اشیاء برجسته تمرکز کنند تا تجربه و تعامل کاربر را با ارائه پوشش ها یا اطلاعات آگاه از زمینه افزایش دهند.

¹ Augmented Reality

² Virtual Reality

1-6- سوالات تحقیق

1. چگونه می‌توان مدل‌های مبدل تصمیم‌گیری را به طور موثر در چارچوب تشخیص شی‌برجسته ادغام کرد تا دقت و استحکام را بهبود بخشد؟
2. مبدل‌های تصمیم از چه مکانیسم‌هایی می‌توانند برای تصمیم‌گیری‌های متوالی برای اصلاح نقشه‌های برجسته و شناسایی اشیاء برجسته در میان صحنه‌های بصری پیچیده استفاده کنند؟
3. چگونه ادغام مدل‌های مبدل تصمیم بر کارایی و سرعت الگوریتم‌های تشخیص اشیاء برجسته، به ویژه برای کاربردهای بلادرنگ تأثیر می‌گذارد؟
4. چالش‌ها و محدودیت‌های کلیدی در ادغام مبدل‌های تصمیم با معماری‌های تشخیص اشیاء برجسته موجود چیست و چگونه می‌توان به این چالش‌ها پرداخت؟

1-7- فرضیات تحقیق

1. ادغام مدل‌های مبدل تصمیم‌گیری در چارچوب‌های تشخیص اشیاء برجسته موجود منجر به پیشرفت‌های قابل‌توجهی در دقت و استحکام می‌شود. استفاده از قابلیت‌های منحصربه‌فرد ترنسفورمرهای تصمیم، مانند توانایی آن‌ها در مدیریت وظایف تصمیم‌گیری متوالی، باعث افزایش دقت در شناسایی مناطق برجسته در تصاویر می‌شود. با این حال، چالش‌های مربوط به یکپارچه‌سازی مدل، مانند سازگاری با معماری‌های موجود و محدودیت‌های منابع محاسباتی، ممکن است ایجاد شود. با توسعه روش‌های مناسب برای پرداختن به این چالش‌ها، ادغام ترنسفورمرهای تصمیم را می‌توان بهینه کرد، که منجر به الگوریتم‌های تشخیص اشیاء برجسته‌تر و قوی‌تر می‌شود.
2. ترنسفورمرهای تصمیم‌گیری می‌توانند مکانیسم‌هایی را برای تصمیم‌گیری متوالی برای اصلاح نقشه‌های برجسته و شناسایی موثر اشیاء برجسته در صحنه‌های بصری پیچیده به کار گیرند. ترنسفورمرهای تصمیم از طریق توانایی پردازش اطلاعات به روشی متوالی، می‌توانند نقشه‌های برجسته را به طور مکرر اصلاح کنند، که منجر به بهبود دقت در شناسایی مناطق مهم بصری می‌شود. با این حال، اثربخشی این مکانیسم‌ها ممکن است تحت تأثیر عواملی مانند پیچیدگی صحنه بصری و تنوع ویژگی‌های شی‌برجسته باشد. با کاوش و تطبیق راهبردهای مختلف تصمیم‌گیری متوالی،

ترنسفورمر های تصمیم می توانند به طور موثر این چالش ها را برطرف کنند و عملکرد الگوریتم های تشخیص اشیاء برجسته را افزایش دهند.

3. ادغام مدل های ترنسفورمر تصمیم گیری در چارچوب های تشخیص شیء برجسته ممکن است بر کارایی و سرعت این الگوریتم ها، به ویژه در برنامه های بلادرنگ تأثیر بگذارد. استفاده از قابلیت های ترنسفورمر های تصمیم گیری برای تصمیم گیری متوالی ممکن است سربار محاسباتی اضافی را معرفی کند که به طور بالقوه بر کارایی و سرعت کلی الگوریتم های تشخیص اشیاء برجسته تأثیر می گذارد. با این حال، از طریق تکنیک های بهینه سازی مانند پردازش موازی، هرس مدل، و شتاب سخت افزاری، می توان تأثیر ترنسفورمر های تصمیم گیری را بر کارایی و سرعت کاهش داد. با ایجاد تعادل بین دقت و کارایی محاسباتی، ترنسفورمر های تصمیم می توانند عملکرد الگوریتم های تشخیص اشیاء برجسته را افزایش دهند و آنها را برای کاربردهای بلادرنگ مناسب کنند.

4. علیرغم مزایای بالقوه ادغام ترنسفورمر های تصمیم با معماری های تشخیص شیء برجسته موجود، چندین چالش و محدودیت کلیدی ممکن است مانع یکپارچه سازی یکپارچه شوند. این چالش ها شامل پیچیدگی مدل، نیازمندی های منابع محاسباتی و سازگاری معماری با چارچوب های موجود است. پرداختن به این چالش ها مستلزم توسعه روش های تخصصی برای انطباق، بهینه سازی و ادغام مدل است. با غلبه بر این موانع، ترنسفورمر های تصمیم می توانند به طور موثر در چارچوب های تشخیص اشیاء برجسته موجود ادغام کرد و در نهایت عملکرد و کاربرد آنها را در مجموعه داده های مختلف و سناریوهای دنیای واقعی افزایش داد.

1-6- ساختار تحقیق

همان طور که در بخش های قبلی بیان شد هدف از انجام این تحقیق تشخیص قسمت های برجسته اشیاء می باشد. در این پژوهش سعی می شود تا ضمن آشنایی، بهترین روش ها نیز معرفی گردد.

در فصل دوم، به معرفی مفاهیم پایه و اصطلاحات استفاده شده، پرداخته می شود. به عبارتی مفاهیمی که در این پژوهش استفاده شده است در این فصل ارائه شده و همچنین به معرفی تشخیص قسمت های برجسته اشیاء و ارزیابی کلی الگوریتم ها پرداخته می شود، که هر یک از روش ها مزایا و معایب خاص خود را دارا می باشند. همچنین در این فصل نیز به مقایسه مطالعات پیشین پرداخته شده است.

در فصل سوم، مدل پیشنهادی اول و دوم شرح داده شده و به توضیح جزئیات معماری هر دو مدل پیشنهادی پرداخته شده است. همچنین در این فصل مقادیر پارامترها نیز بیان شده است.

در فصل چهارم مدل پیشنهادی ارزیابی شده و با روش‌های موجود مقایسه می‌شود. ارزیابی سیستم با سه آزمایش، اول انتخاب بهترین مقادیر پارامترها، دوم مقایسه دو مدل پیشنهادی و سوم مقایسه مدل پیشنهادی با سایر روش‌ها انجام شده و نتایج به‌دست آمده در جداول و نمودارهای این فصل ارائه شده است. در نهایت نتایج مدل پیشنهادی با نتایج سایر مطالعات مقایسه می‌شود و در نهایت در فصل پنجم، نتایج به‌دست آمده و کارهای آینده بیان می‌شود.

فصل 2: ادبیات نظری و پیشینه تحقیق

2-1- مقدمه

تشخیص قسمت های برجسته ی شی¹، وظیفه شناسایی و بخش‌بندی مناطق مهم بصری در تصاویر، به عنوان یک منطقه تحقیقاتی حیاتی در بینایی کامپیوتر ظاهر شده است. توانایی تشخیص و بومی‌سازی خودکار این مناطق، تمرکز ذاتی سیستم بینایی انسان را بر مناطق مورد علاقه منعکس می‌کند و طیف وسیعی از برنامه‌های کاربردی را در بینایی رایانه، پردازش چند رسانه‌ای و تعامل انسان و رایانه امکان‌پذیر می‌سازد. پیشرفت‌های اخیر در یادگیری عمیق [28]، به‌ویژه پذیرش شبکه‌های عصبی کاملاً کانولوشن² و دیگر معماری‌های عمیق، دقت و استحکام روش‌های تشخیص قسمت های برجسته ی شی را به طور قابل‌توجهی افزایش داده است.

رویکردهای متعارف و مبتنی بر یادگیری عمیق برای تشخیص قسمت های برجسته ی شی به چالش‌های متعددی مانند پس‌زمینه‌های پیچیده، مقیاس‌های مختلف شی و سوگیری‌های مجموعه داده پرداخته است. این روش‌ها از ویژگی‌های مهندسی، الگوهای یادگیری و نوآوری‌های معماری برای بهبود عملکرد استفاده می‌کنند. بررسی‌های جامع این زمینه بر تکامل تکنیک‌ها، معیارهای ارزیابی به کار گرفته شده و معیارهای مورد استفاده برای تجزیه و تحلیل مقایسه‌ای تاکید کرده است. با این حال، چالش‌هایی مانند هزینه‌های محاسباتی بالا، محدودیت‌های حافظه، تعمیم به سناریوهای مختلف، و انعطاف‌پذیری در برابر حملات متخاصم همچنان موانعی حیاتی برای استفاده در دنیای واقعی روش‌های اخیر هستند.

در حوزه تصاویر سنجش از دور³، تأکید بر معماری‌های سبک وزن و در عین حال مؤثر، نیاز به راه‌حلی را برجسته می‌کند که دقت تشخیص را با کارایی محاسباتی متعادل کند. نوآوری‌هایی مانند مکانیسم‌های توجه، هدایت معنایی و پردازش چند مقیاسی در پرداختن به چالش‌های خاص تشخیص قسمت های برجسته ی شی، به‌ویژه برای محیط‌های محدود به منابع، نویدبخش بوده است.

این پایان‌نامه یک رویکرد جدید برای تشخیص قسمت های برجسته اشیاء با ادغام مبدل‌های تصمیم، کلاسی از مدل‌ها که به دلیل قابلیت‌های الهام‌گرفته از یادگیری تقویتی مشهور هستند، در چارچوب تشخیص قسمت های برجسته ی شی معرفی می‌کند. مبدل‌های تصمیم، با توانایی خود در ترتیب دادن وظایف و استفاده مؤثر از داده‌های تاریخی، یک تغییر پارادایم قانع‌کننده برای مدل‌سازی برجسته ارائه می‌دهند.

¹ Saliency Object Detection

² Fully Connected Layer

³ Remote Sensing Image

هدف پایان نامه با ترکیب این معماری، پرداختن به چالش های پایدار در تشخیص قسمت های برجسته ی شی، از جمله استحکام در شرایط متخاصم، مقیاس پذیری در مجموعه داده های متنوع، و ادغام اطلاعات متنی و زمانی برای عملکرد بهبود یافته است.

روش پیشنهادی با معیارهای پیشرفته با استفاده از معیارهای تعیین شده برای دقت تشخیص، کارایی محاسباتی و استحکام ارزیابی خواهد شد. از طریق این کار، هدف ما ارائه دیدگاهی دگرگون کننده به حوزه تشخیص اشیاء برجسته است، و پیشرفت هایی را تقویت می کند که هم با نوآوری نظری و هم با کاربرد عملی هماهنگ است. از این رو در این فصل ابتدا تعریف قسمت های برجسته ی اشیاء و سپس انواع روش های تشخیص قسمت های برجسته ی اشیاء مورد بررسی قرار گرفته است.

2-2- تشخیص قسمت های برجسته ی اشیاء

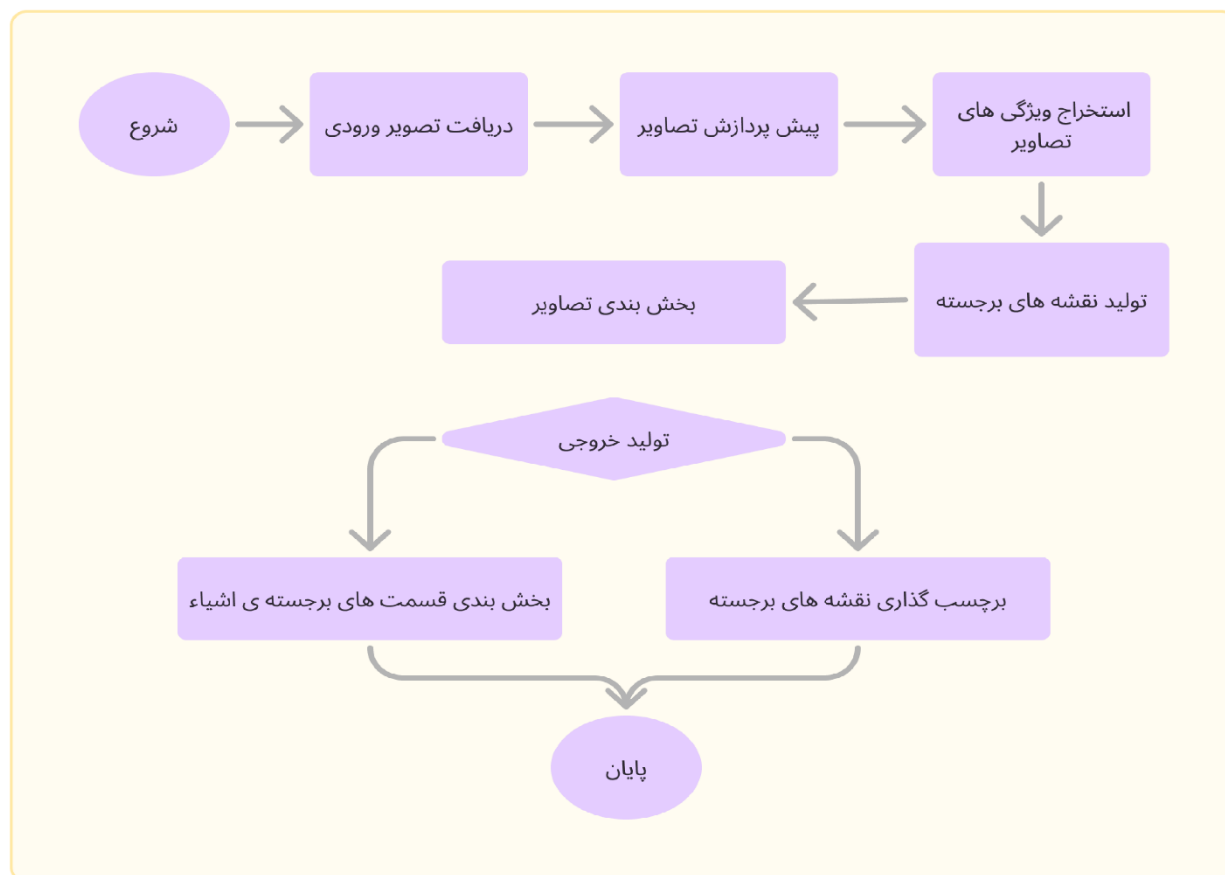
تشخیص قسمت های برجسته ی اشیاء زیرشاخه ای از بینایی کامپیوتری است که بر شناسایی و تقسیم بندی مهم ترین مناطق بصری در یک تصویر تمرکز دارد. این مناطق که از آنها به عنوان اشیاء برجسته یاد می شود، به دلیل ویژگی های متمایز خود در مقایسه با مناطق اطراف، طبیعتاً توجه انسان را به خود جلب می کنند.

قسمت های برجسته اشیاء مناطق خاصی در یک تصویر هستند که به دلیل ویژگی هایی مانند رنگ، بافت، روشنایی، کنتراست یا ساختار هندسی برجسته می شوند. این ویژگی ها شی را از پس زمینه یا مناطق کم اهمیت آن متمایز می کند و آنها را در کانون توجه قرار می دهد. به عنوان مثال، یک گل روشن در یک جنگل تاریک یا یک سیب قرمز در میان برگ های سبز نمونه هایی از قسمت های برجسته هستند که تمرکز بصری را جلب می کنند.

2-3- سیستم تشخیص قسمت های برجسته ی اشیاء

همان طور که گفته شد، در تشخیص قسمت های برجسته ی اشیاء تمرکز بر روی شناسایی و بخش بندی خودکار مهم ترین مناطق از نظر بصری است. این سیستم بر مناطقی که به طور طبیعی توجه را به خود جلب میکنند، تمرکز دارد، مانند اشیاء یا ویژگی هایی که از محیط اطراف خود متمایز هستند، از سیستم بینایی انسان تقلید می کند.

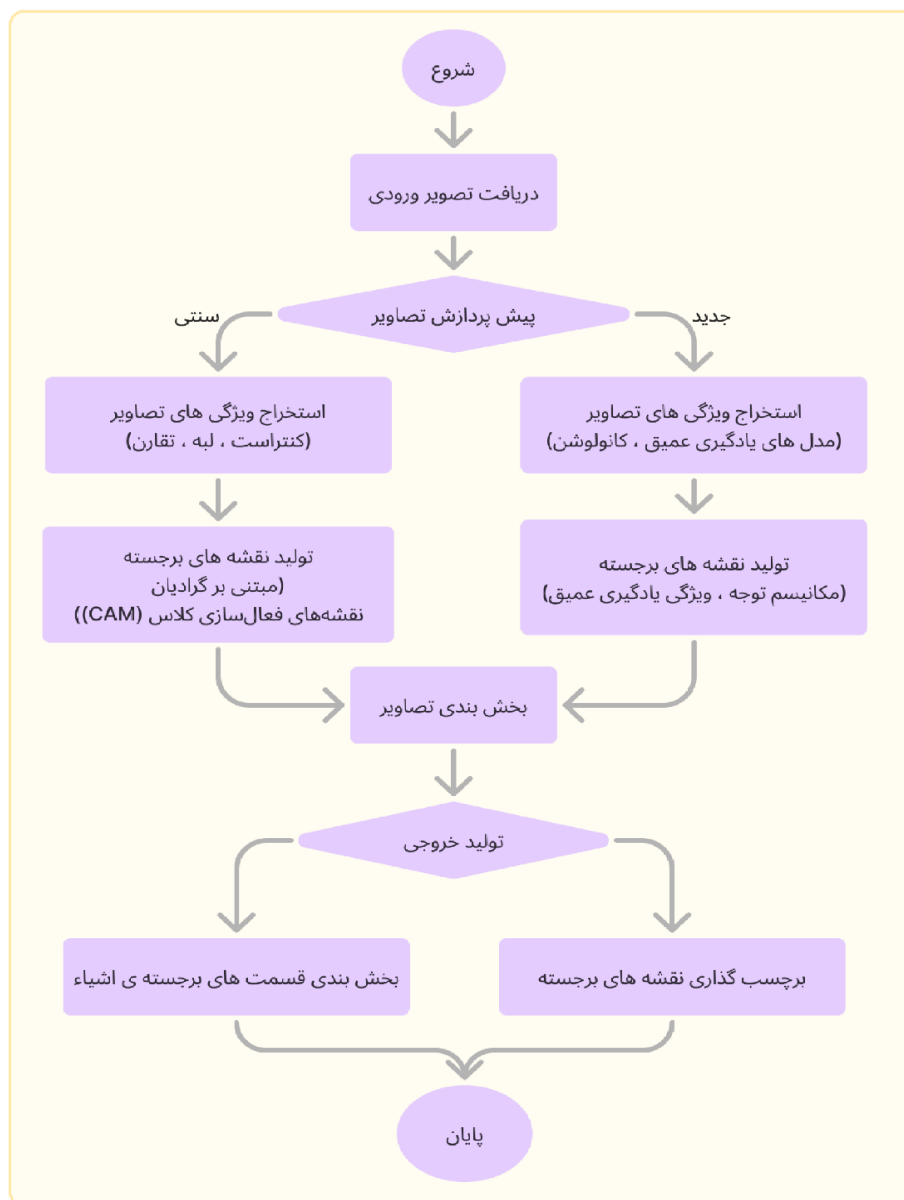
عملکرد اصلی سیستم تجزیه و تحلیل یک تصویر و ایجاد یک نقشه برجسته است، نمایی که در آن مناطق با شدت بالا با اشیاء برجسته مطابقت دارند. سپس از این نقشه برای برجسته کردن یا جداسازی این اشیاء برای پردازش بیشتر استفاده می شود.



شکل (1-2) تشخیص قسمت های برجسته ی شی

شکل (1-2) به طور کلی روش های تشخیص قسمت های برجسته را پوشش می دهد با این تفاوت که روش های سنتی تشخیص برجستگی بر ویژگی های دست ساز مانند کنتراست رنگ، الگوهای بافت، و شیب لبه ها، با استفاده از الگوریتم های ساده که در سناریوهای کنترل شده به خوبی عمل می کردند، اما با صحنه های پیچیده یا ظاهر اشیاء متفاوت دست و پنجه نرم می کردند، متکی بودند. در مقابل، رویکردهای یادگیری عمیق مدرن از شبکه های عصبی برای استخراج خودکار ویژگی های سلسله مراتبی استفاده می کند و عملکرد قوی را در تصاویر چالش برانگیز با انسداد و پس زمینه های به هم ریخته ممکن می سازد. مدل های سنتی، اگرچه از نظر محاسباتی کارآمد و قابل تفسیر هستند، اما فاقد مقیاس پذیری و سازگاری بودند و اغلب به تنظیم دستی نیاز داشتند. روش های پیشرفته امروزی، مانند شبکه ها و مدل های کاملاً کانولوشن،

از مجموعه داده‌های مقیاس بزرگ، مکانیسم‌های توجه و مدل‌های از پیش آموزش‌دیده برای دستیابی به دقت و استحکام بالا، حتی در شرایط متخاصم، استفاده می‌کنند. معماری‌های سبک وزن همچنین محدودیت‌های منابع را برطرف می‌کنند و این سیستم‌های پیشرفته را در بین برنامه‌های مختلف متنوع‌تر می‌کنند. در حالی که روش‌های مدرن در مقیاس‌پذیری و تعمیم برتری دارند، اتکای آن‌ها به مجموعه داده‌های بزرگ و منابع محاسباتی یک مبادله را در مقایسه با تکنیک‌های سنتی ساده‌تر برجسته می‌کند. در شکل (2-2) می‌توانیم تفاوت دو روش را به وضوح مشاهده کرد.



شکل (2-2) تفاوت تشخیص قسمت های برجسته ی شی روش های سنتی و مدرن

2-4- روش های تشخیص قسمت های برجسته ی اشیاء

تشخیص قسمت های برجسته ی شی با توسعه تکنیک های مختلف مرتبط [24]، از جمله پیش بینی تثبیت، تشخیص اشیاء برجسته، تشخیص همبرجستگی، و تشخیص برجسته بودن عمق و رنگی بودن تصاویر¹ [29]، پیشرفت قابل توجهی داشته است. هدف مدل های پیش بینی تثبیت، پیش بینی حرکات چشم انسان و مناطقی از تصویر است که توجه را به خود جلب می کند و بینش های ارزشمندی را در مورد برجستگی بصری ارائه می دهد. با تکیه بر این، تشخیص شی برجسته بر شناسایی و بخش بندی برجسته ترین شی (های) بصری در یک تصویر تمرکز می کند، که اغلب از نقشه های ثابت برای هدایت فرآیند تشخیص استفاده می کند. تشخیص مشارکتی با تشخیص اشیاء برجسته که در چندین تصویر مشترک هستند، این را بیشتر گسترش می دهد، که به ویژه در سناریوهای تشخیص چند شی مفید است. تشخیص برجسته بودن عمق و رنگی بودن تصاویر هم عمق و هم اطلاعات رنگی را برای شناسایی اشیاء برجسته در فضاهای سه بعدی ترکیب می کند [29] و لایه دیگری از پیچیدگی را به فرآیند تشخیص اضافه می کند. رویکرد مرتبط دیگر، پیش بینی نگاه اجتماعی، تمرکز توجه را در صحنه های گروهی مطالعه می کند، الگوهای نگاه جمعی افراد [36] را پیش بینی می کند، بنابراین رفتار اجتماعی را با برجستگی بصری پیوند می دهد. این تکنیک ها، اگرچه همگی مربوط به برجسته بودن هستند، اما در کاربردها و رویکردهایشان متفاوت هستند. با حرکت رو به جلو، روش های تشخیص قسمت های برجسته ی شی را می توان به طور کلی به دو دسته طبقه بندی کرد: روش های سنتی [1,3,4,5,6,23,24,27]، که بر ویژگی های دست ساز و مدل های اکتشافی تکیه دارند، و روش های مبتنی بر یادگیری عمیق [7,11,12,13,15,19,37,47]، که از شبکه های عصبی پیشرفته برای یادگیری خودکار الگوهای پیچیده از مجموعه داده های در مقیاس بزرگ استفاده می کنند. تغییر به سمت یادگیری عمیق دقت و استحکام تشخیص برجستگی را به ویژه در محیط های چالش برانگیز به طور قابل توجهی افزایش داده و مدل های بیولوژیکی قابل قبول تری را الهام گرفته است که با فرآیندهای توجه انسان همسو می شوند.

¹ Red, Green, Blue – Depth (RGB-D)

2-4-1- پیش بینی تمرکز دید¹

پیش‌بینی تمرکز دید تکنیکی است که برای پیش‌بینی جایی که ناظر انسانی احتمالاً نگاه خود را در یک تصویر یا صحنه متمرکز می‌کند استفاده می‌شود. بر اساس عواملی مانند برجسته بودن تصویر و دانش قبلی از رفتار بصری انسان، حرکات چشم را شبیه سازی می‌کند و مناطقی را که توجه بصری را به خود جلب می‌کنند، پیش بینی می‌کند. این اغلب از طریق مدل‌های محاسباتی به دست می‌آید که تمرکز سیستم بینایی انسان را بر روی ویژگی‌های خاص (مانند کنتراست، حرکت) در یک صحنه تقلید می‌کنند. معمولاً در تعامل انسان و رایانه، مدل‌سازی توجه بصری، و وظایف مبتنی بر برجسته‌سازی مانند تشخیص اشیا استفاده می‌شود.

2-4-2- تشخیص هم برجستگی²

تشخیص هم برجستگی توسعه ای از تشخیص قسمت های برجسته ی شی است که هدف آن شناسایی اشیاء برجسته رایج در مجموعه ای از تصاویر است. اشیایی را شناسایی می‌کند که به طور مشابه در چندین تصویر برجسته هستند، که اغلب در زمینه شناسایی اشیایی که دارای ویژگی های مشترک در چندین نما یا نمونه های مختلف هستند استفاده می‌شود. تشخیص مشترک برجستگی هنگام رسیدگی به سناریوهایی مانند کشف شی یا تشخیص چند شی در تصاویر مرتبط مهم است. این تکنیک معمولاً در ردیابی چند شیء، تجزیه و تحلیل عکس گروهی و درک صحنه استفاده می‌شود.

2-4-3- تشخیص عمق برجستگی تصاویر رنگی³

تشخیص عمق برجستگی تصاویر رنگی شامل استفاده از اطلاعات رنگ⁴ و عمق⁵ برای تشخیص اشیاء برجسته در یک فضای سه بعدی است. در حالی که رنگ و بافت اشیاء را ثبت می‌کند، اطلاعات عمق بعد سوم را اضافه می‌کند و زمینه فضایی را فراهم می‌کند. این ترکیب به مدل اجازه می‌دهد تا اشیایی را که نه تنها بر اساس ویژگی‌های بصری بلکه بر اساس ارتباط فضایی آنها در صحنه‌های سه‌بعدی برجسته

¹ Fixation Prediction

² Co-Saliency Detection

³ RGB-D Saliency Detection

⁴ RGB

⁵ Depth

هستند، شناسایی کند. اغلب در رباتیک، واقعیت افزوده¹ و تشخیص اشیاء سه بعدی، که در آن اطلاعات عمق برای درک محیط بسیار مهم است، استفاده می شود.

2-4-4- پیش بینی نگاه اجتماعی²

پیش بینی نگاه اجتماعی به پیش بینی تمرکز توجه در یک صحنه گروهی با در نظر گرفتن الگوهای نگاه افراد درون گروه اشاره دارد. این برجستگی بصری را با رفتار اجتماعی با پیش بینی اینکه افراد یک گروه احتمالاً به کجا نگاه می کنند، بر اساس نشانه های اجتماعی مشترک و الگوهای نگاه، پیوند می دهد. این تکنیک نحوه تأثیرگذاری پویایی های اجتماعی، مانند رفتار و تعامل گروهی بر تمرکز توجه را مدل می کند. در تعامل انسان و ربات، تجزیه و تحلیل جمعیت، و درک محتوای رسانه های اجتماعی استفاده می شود، جایی که درک تمرکز جمعی می تواند تجربه کاربر را افزایش دهد یا بینش هایی را در مورد رفتار گروه ارائه دهد.

2-4-5- تشخیص شیء برجسته

تشخیص شیء برجسته به وظیفه شناسایی و تقسیم بندی برجسته ترین اشیاء بصری در یک تصویر اشاره دارد. این اشیاء برجسته در نظر گرفته می شوند زیرا به دلیل رنگ، کنتراست یا سایر ویژگی های ادراکی خود برجسته می شوند. مدل های تشخیص قسمت های برجسته ی شیء را می توان برای کارهایی مانند تقسیم بندی تصویر، محلی سازی شیء، و برجسته کردن مناطق مهم برای تجزیه و تحلیل بیشتر استفاده کرد. در بینایی کامپیوتر برای حاشیه نویسی خودکار تصویر، تشخیص اشیاء و دستکاری تصویر آگاهانه از محتوا بسیار مهم است.

2-4-6- مقایسه روش ها

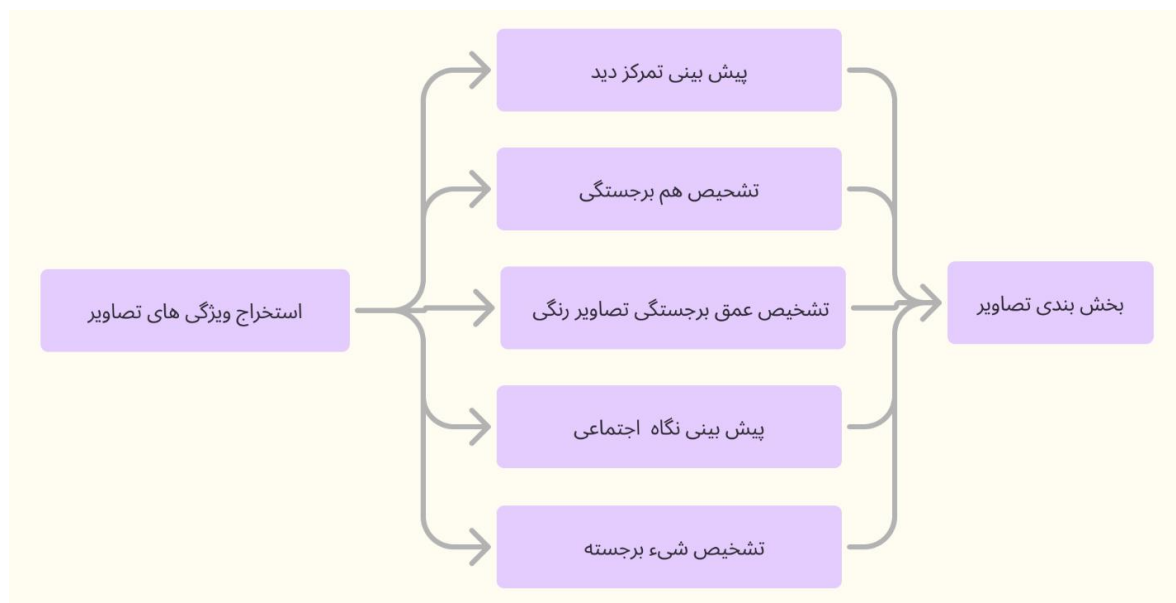
مقایسه روش های مختلف در تشخیص شیء برجسته نشان می دهد که هر کدام از این تکنیک ها هدف خاصی دارند و برای کاربردهای مختلف طراحی شده اند. پیش بینی تمرکز دید بر شبیه سازی حرکت چشم انسان و پیش بینی مناطقی از تصویر که بیشترین توجه را جلب می کنند تمرکز دارد [22] و بیشتر در زمینه های تعامل انسان با کامپیوتر و مدل سازی توجه انسانی استفاده می شود [43]. در حالی که تشخیص قسمت های برجسته ی شیء هدفش شناسایی و بخش بندی اشیاء بصری برجسته در یک تصویر است [4,7,37] و در

¹ Augmented Reality

² Social Gaze Prediction

کاربردهایی مانند شناسایی اشیاء و دستکاری خودکار تصاویر استفاده می‌شود. تشخیص هم‌برجستگی¹ که نسخه‌ای گسترش یافته از تشخیص قسمت های برجسته ی شی است [17, 36]، اشیاء برجسته مشترک در مجموعه‌ای از تصاویر را شناسایی می‌کند و در زمینه‌هایی مانند شناسایی اشیاء چندگانه یا تحلیل عکس‌های گروهی به کار می‌رود. تشخیص تشخیص برجسته بودن عمق و رنگی بودن تصاویر [29, 27] ترکیبی از اطلاعات رنگی و عمقی است که در فضاهای سه‌بعدی برای شناسایی اشیاء برجسته استفاده می‌شود و در کاربردهایی مانند رباتیک و واقعیت افزوده اهمیت دارد. در نهایت، پیش‌بینی نگاه اجتماعی به پیش‌بینی تمرکز توجه در صحنه‌های گروهی می‌پردازد [36] و بر الگوهای نگاه اجتماعی تأکید دارد، که در تحلیل رفتار گروهی و تعاملات اجتماعی مفید است. این روش‌ها هرکدام با استفاده از ویژگی‌های خاص خود، به شکلی متفاوت به مسائل مرتبط با برجستگی دیداری پرداخته و در حوزه‌های مختلفی از بینایی ماشین و تحلیل اجتماعی کاربرد دارند.

شکل (2-3) مقایسه ی رویکرد های مشابه تشخیص قسمت های برجسته اشیاء



¹ Co-Saliency Detection

2-5- مدل‌های سنتی تشخیص قسمت های برجسته اشیاء

2-5-1- روش های مبتنی بر کنتراست¹

روش‌های مبتنی بر کنتراست جهانی [1] بر تفاوت ظاهری بین یک شی و پس‌زمینه آن در کل تصویر تمرکز می‌کنند. ایده این است که اشیاء برجسته معمولاً با تفاوت های جهانی شدید در ویژگی هایی مانند رنگ، شدت یا بافت از پس زمینه قابل تشخیص هستند. تشخیص منطقه برجسته مبتنی بر کنتراست جهانی نمونه ای [1] از این رویکرد است، که در آن تصویر برای یافتن مناطقی با بالاترین کنتراست که می تواند با اشیاء برجسته مطابقت داشته باشد، تجزیه و تحلیل می شود. رویکردهای کنتراست محلی بر تفاوت‌های کنتراست پیکسلی یا ناحیه‌ای در محله‌های محلی تمرکز می‌کنند و درک دقیق‌تری از ساختار بصری تصویر ارائه می‌دهند. کنتراست محلی اغلب همراه با کنتراست جهانی برای بهبود تشخیص برجسته بودن با پرداختن به جزئیات دقیق‌تر و وابسته به زمینه که کنتراست جهانی ممکن است نادیده گرفته شود، استفاده می‌شود.

2-5-2- روش های منطقه محور²

این روش ها از ویژگی های منطقه ای برای تشخیص اشیاء برجسته از پس زمینه استفاده می کنند. با تجزیه و تحلیل مناطق مختلف در یک تصویر، این روش‌ها ویژگی‌های محلی را به تصویر می‌کشند که یک شی را برجسته می‌کند. رویکرد یکپارچه سازی ویژگی های منطقه ای متمایز از یک رویکرد متمایز برای ادغام ویژگی های منطقه ای استفاده می کند و شناسایی اشیاء برجسته را بهبود می بخشد. روش‌های مبتنی بر منطقه سلسله مراتبی تصویر را در مقیاس‌های چندگانه یا سطوح انتزاعی تجزیه و تحلیل می‌کنند و به سیستم اجازه می‌دهند مناطق برجسته را در سطوح مختلف دانه‌بندی شناسایی کند. این رویکرد به شناسایی اجسام بزرگ و کوچک و همچنین جزئیات دقیق کمک می کند. روش تشخیص برجستگی سلسله مراتبی نمونه ای از این است که مناطق تصویر را در مقیاس های مختلف پردازش می کند تا برجسته بودن اشیا را بهتر به تصویر بکشد.

¹ Contrast-Based Method

² Region-Based Method

2-5-3- روش های مبتنی بر پیشینه ی پیشین¹

این روش ها با فرض اینکه پس زمینه معمولاً توجه بصری کمتری دارد، از اطلاعات پس زمینه برای اصلاح تشخیص برجسته سازی استفاده می کنند. این روش ها با تأکید بر اشیاء پیش زمینه که با پس زمینه تضاد دارند، نقشه های برجسته را بهبود می بخشند. تکنیک برجستگی ژئودزیک با استفاده از پیش زمینه های پس زمینه² مثالی است که از پیش زمینه های پس زمینه برای تعیین اهمیت پیکسل ها نسبت به پس زمینه استفاده می کند. روش های مبتنی بر شیء قبلی، دانش قبلی را در مورد شکل، اندازه یا موقعیت های معمولی اشیاء در یک صحنه به منظور بهبود تشخیص برجسته بودن ترکیب می کنند. این پیشین ها، مدل را به سمت تمرکز بر مناطقی هدایت می کنند که احتمالاً حاوی اشیاء مهم هستند. این روش ها با در نظر گرفتن ظاهر مورد انتظار اشیاء، نقشه های برجسته را افزایش می دهند و آنها را در تشخیص ویژگی های بصری مرتبط مؤثرتر می سازند.

2-5-4- روش های مبتنی بر انتقال³

این تکنیک [4] اطلاعات برجسته را از تصاویر مرجع یا مجموعه داده ها به تصاویر هدف منتقل می کند و اغلب از تطابق بین ویژگی های تصویر استفاده می کند. روش مکاتبه محور انتقال برجسته⁴ نمونه ای است که در آن اطلاعات برجسته از یک تصویر به تصویر دیگر بر اساس ویژگی های بصری مشترک یا مناطق مربوطه نگاشت می شود و به شناسایی اشیاء برجسته در تصاویر جدید با استفاده از الگوهای آموخته شده کمک می کند.

2-5-5- روش های مبتنی بر بهینه سازی⁵

روش های مبتنی بر بهینه سازی تشخیص برجستگی را به عنوان یک مسئله بهینه سازی فرموله می کنند، جایی که هدف اصلاح نقشه های برجستگی با به حداقل رساندن توابع انرژی است. هدف این توابع افزایش تشخیص اشیاء برجسته در عین به حداقل رساندن موارد مثبت کاذب است. روش بهینه سازی برجسته از تشخیص پس زمینه قوی نمونه ای است که در آن نقشه های برجسته به طور مکرر برای بهبود استحکام

¹ Prior-Based Method

² Geodesic Saliency Using Background Priors

³ Transfer-Based Method

⁴ Correspondence Driven Saliency Transfer

⁵ Optimization-Based Method

تشخیص پس زمینه اصلاح می شوند. رویکردهای بهینه سازی مبتنی بر نمودار، تشخیص برجسته سازی را به عنوان یک مشکل گراف در نظر می گیرند، جایی که گره ها پیکسل های تصویر یا مناطق را نشان می دهند و لبه ها روابط آنها را نشان می دهند. اطلاعات برجسته از طریق نمودار منتشر می شود تا نقشه برجستگی را اصلاح کند. این روش ها در تشخیص اشیا و جداسازی آنها از پس زمینه با مدل سازی روابط فضایی بین پیکسل ها موثر هستند.

2-5-6- روش های هندسی و مبتنی بر فاصله¹

روش های فاصله ژئودزیکی [3] فاصله بین پیکسل ها یا مناطق را برای شناسایی برجستگی اندازه گیری می کنند. با در نظر گرفتن رابطه بین مرزهای شی و مناطق پس زمینه، این روش ها اشیاء برجسته را با تمرکز بر مناطقی با حداقل فاصله ژئودزیکی تا پس زمینه شناسایی می کنند و عناصر بصری مهم را برجسته می کنند. روش برجستگی ژئودزیکی [3] با استفاده از پیش زمینه های پس زمینه² از این مفهوم برای اصلاح نقشه های برجسته استفاده می کند و بر جداسازی شی از پس زمینه تمرکز دارد.

2-5-7- روش های ترکیبی³

روش های ترکیبی اصول متعددی را از دسته های مختلف ترکیب می کنند مانند ویژگی های مبتنی بر کنتراست، تجزیه و تحلیل مبتنی بر منطقه و دانش قبلی برای افزایش عملکرد تشخیص برجسته. با استفاده از تکنیک های مختلف، این روش ها می توانند مجموعه ای جامع تر از نشانه های بصری را ثبت کنند و دقت و استحکام نقشه های برجسته را بهبود بخشند. بسیاری از رویکردهای سنتی مدرن چنین روش های ترکیبی را برای رفع محدودیت های تکنیک های فردی و دستیابی به نتایج تشخیص بهتر در صحنه های پیچیده ادغام می کنند.

¹ Geometric and Distance Based Method

² Geodesic Saliency Using Background Priors

³ Hybrid Method

2-6- مدل‌های یادگیری عمیق تشخیص قسمت های برجسته اشیاء

2-6-1- شبکه های عصبی کانولوشنی¹

شبکه های کاملاً کانولوشنی²، معماری های تخصصی شبکه های کانولوشنی هستند که در آن همه لایه ها کانولوشن هستند و آنها را برای تشخیص برجسته بودن در سطح پیکسل مناسب می کند. آنها به ویژه برای تقسیم بندی اشیاء برجسته از اطلاعات پس زمینه در فیلم ها و تصاویر، همانطور که در "تشخیص اشیاء برجسته ویدئویی از طریق شبکه های کاملاً پیچیده" [7]، موثر هستند. ویژگی های کانولوشنال نامشخص عدم قطعیت را در ویژگی های کانولوشن معرفی می کند و به شبکه اجازه می دهد تا نقشه های برجسته را اصلاح کند و دقت تشخیص اشیاء را بهبود بخشد. با یادگیری ویژگی های نامطمئن [18]، روش نشان داده شده [18] می تواند ابهامات را کنترل کند و پیش بینی های برجسته را بهبود بخشد. هدف یادگیری کنتراست عمیق یادگیری ویژگی های عمیقی است که بر کنتراست تأکید دارند و به تشخیص بهتر شیء برجسته از محیط اطراف کمک می کنند [11]. کار [7,10,13] با یادگیری ویژگی هایی که تضاد بین اشیاء مهم و پس زمینه را برجسته می کند، فرآیند تشخیص برجسته بودن را افزایش می دهد.

2-6-2- شبکه های عصبی بازگشتی³

این روش از شبکه های عصبی بازگشتی همراه با مکانیسم های توجه [9] برای تمرکز مکرر بر برجسته ترین مناطق یک تصویر یا ویدئو استفاده می کند. رویکرد، نشان داده [18,11]، به مدل اجازه می دهد تا پیش بینی های خود را با بازدید مجدد از مناطق مورد نظر اصلاح کند و دقت تشخیص برجسته بودن را بهبود بخشد. شبکه کاملاً کانولوشنی مبتنی بر منطقه⁴ مزایای شبکه های عصبی مکرر را با لایه های کاملاً کانولوشن ترکیب می کنند و مدل را قادر می سازند تا نقشه های برجسته را به طور مکرر اصلاح کنند. [10,13] این معماری به افزایش تدریجی نقشه برجسته با یادگیری اطلاعات زمینه ای بیشتر در حین تکرار کمک می کند.

¹ Convolutional Neural Network

² Fully Convolutional Neural Network

³ Recurrent Neural Network

⁴ Region-based Convolutional Network

2-6-3- مدل های عمیق سلسله مراتبی¹

شبکه سلسله مراتبی عمیق از یک معماری سلسله مراتبی برای استخراج ویژگی های برجسته چند سطحی از یک تصویر استفاده می کند. با پردازش تصاویر در سطوح مختلف انتزاع [7]، تشخیص برجسته تری را در مقیاس های مختلف ویژگی های بصری امکان پذیر می کند. مدل سازی توجه برای محل سازی شی جهانی² ویژگی های کانولوشن چند سطحی را جمع آوری می کند تا تشخیص برجسته سازی را افزایش دهد. با ترکیب ویژگی های سطوح مختلف شبکه [7,10,13]، این مدل با ثبت جزئیات دقیق و نمایش های سطح بالاتر، نقشه های برجسته تری را تولید می کند.

2-6-4- مکانسیم های مبتنی بر توجه³

این مدل ها مکانسیم های توجه را به طور مکرر اعمال می کنند تا بر برجسته ترین مناطق در یک دنباله از فریم ها یا یک تصویر تمرکز کنند. مدل های توجه مکرر، مانند مدل پیشنهادی [9] پیش بینی های برجستگی را با بازدید مجدد چندین نوبت از امیدوار کننده ترین مناطق مورد علاقه اصلاح کنید. این رویکرد به تدریج نقشه های برجسته را با استفاده از مکانسیم های توجه در مراحل مختلف پردازش اصلاح می کند. روش نشان داده شده [9,18]، به طور مکرر پیش بینی های برجسته را تنظیم می کند و به مدل اجازه می دهد در هر مرحله به ویژگی های مهم یک تصویر یا ویدیو توجه بیشتری داشته باشد.

2-6-5- یادگیری تحت نظارت با ویژگی های عمیق⁴

شبکه های تحت نظارت عمیق، نظارت را در لایه های متعدد درون شبکه ترکیب می کنند، که امکان یادگیری بهتر ویژگی های برجسته را در سراسر عمق شبکه فراهم می کند. این تکنیک، [7,13] فرآیند تشخیص برجسته بودن را با اعمال نظارت در لایه های میانی برای هدایت فرآیند یادگیری بهبود می بخشد. مجموعه های سطح عمیق از شبکه های عصبی عمیق برای انجام بخش بندی بر اساس مجموعه سطح اشیاء برجسته استفاده می کنند. با استفاده از ویژگی های عمیق در تقسیم بندی سطح مجموعه، [11] این رویکرد دقت مرزهای شی شناسایی شده را بهبود می بخشد و منجر به نقشه های برجسته بهتر می شود.

¹ Deep Hierarchical Spatial Network

² Attention Modeling for Universal Object Localization

³ Attention-Based Mechanisms

⁴ Supervised Learning with Deep Features

2-6-6- ویژگی های متنی و جهانی¹

این روش اطلاعات زمینه جهانی و محلی را در فرآیند تشخیص برجسته بودن ادغام می کند. با در نظر گرفتن وابستگی های زمینه ای، [9,10,13] چند زمینه ای توانایی مدل را برای تشخیص اشیاء برجسته بهبود می بخشد و آن را در محیط های متنوع قوی تر می کند. ویژگی های عمقی غیرمحلی [18] وابستگی های دوربرد بین پیکسل ها یا مناطق یک تصویر را ثبت می کنند. این رویکرد، [18] تشخیص برجسته بودن را با در نظر گرفتن اطلاعات جهانی افزایش می دهد و به مدل اجازه می دهد اشیاء برجسته را بر اساس نشانه های زمینه ای گسترده تر شناسایی کند.

2-6-7- ویژگی های چند مقیاسه²

این تکنیک شامل استخراج ویژگی ها از مقیاس های متعدد برای بهبود استحکام و دقت تشخیص برجسته است. با در نظر گرفتن جزئیات دقیق و ویژگی های زمینه ای گسترده تر، مدل، [13] بهتر می تواند اشیاء برجسته را در وضوح های فضایی مختلف تشخیص دهد. ویژگی های چند مقیاسی رمزگذاری شده، نقشه های فاصله سطح پایین را با ویژگی های سطح بالا از مقیاس های مختلف ترکیب می کنند تا تشخیص برجسته سازی را بهبود بخشند [7,18]. این روش ها، [7,10,13,18] نقشه برجسته را با استفاده از اطلاعات دقیق و جهانی به شیوه ای یکپارچه افزایش می دهد.

2-6-8- رویکرد های ترکیبی و بهینه سازی³

این مدل ها تخمین محلی را با تکنیک های جستجوی جهانی ترکیب می کنند تا تشخیص برجستگی را بهبود بخشند. با استفاده از نشانه های محلی برای تشخیص اولیه و جستجوی جهانی برای اصلاح، [9,11] این رویکردها با در نظر گرفتن اطلاعات محلی و جهانی در فرآیند برجسته سازی، دقت را بهبود می بخشند.

2-7- مدل های مبدل تصمیم⁴

مدل های مبدل تصمیم یک نوآوری اخیر در زمینه تصمیم گیری متوالی است که نقاط قوت یادگیری تقویتی و مدل سازی توالی عمیق را با استفاده از مبدل ها ادغام می کند. بر خلاف روش های یادگیری تقویتی سنتی،

¹ Contextual and Global Features

² Multi-Scale Features

³ Hybrid and Optimization Approaches

⁴ Decision Transformer Model

که بر توابع ارزش یا رویکردهای گرادیان خط مشی تکیه می‌کنند، مبدل‌های تصمیم با در نظر گرفتن آن به عنوان یک کار مدل‌سازی دنباله‌ای نظارت شده [20,21]، به مشکل نزدیک می‌شوند.

این مدل‌ها با استفاده از معماری الهام‌گرفته از مبدل‌های تصمیم، که در ابتدا برای کارهای پردازش زبان طبیعی مانند ترجمه ماشینی توسعه داده شده‌اند، مسیرها - توالی حالت‌ها، اقدامات و پاداش‌ها را پردازش می‌کنند. با استفاده از قدرت مکانیزم توجه [20,22]، مبدل تصمیم می‌تواند وابستگی‌های دوربرد را در مسیرها مدل‌سازی کند و آنها را قادر می‌سازد تا سیاست‌های بهینه را مستقیماً از داده‌های تاریخی بیاموزند.

یکی از پیشرفت‌های کلیدی مبدل‌های تصمیم، استفاده از بازگشت به رفتن¹ به عنوان یک سیگنال راهنما است. آنها به جای پیش‌بینی ارزش اقدامات یا به حداکثر رساندن پاداش‌های فوری، پیش‌بینی‌ها را به پاداش جمعی مورد نظر در آینده مشروط می‌کنند. این باعث می‌شود مدل بسیار انعطاف‌پذیر باشد و به آن اجازه می‌دهد تا با وظایف مختلف با ساختارهای پاداش متفاوت سازگار شود [18,20].

تشخیص قسمت‌های برجسته ی اشیاء با شناسایی و برجسته کردن برجسته ترین اشیاء در یک تصویر، نقش مهمی در درک تصویر و وظایف بینایی رایانه ایفا می‌کند. کاربردهای آن در حوزه‌های مختلفی مانند تقسیم‌بندی تصویر، ردیابی شی، و تجزیه و تحلیل ویدئو، که در آن به تمرکز توجه بر مرتبط‌ترین بخش‌های یک صحنه کمک می‌کند، می‌شود. در زمینه تشخیص قسمت‌های برجسته ی اشیاء، یک چارچوب یکپارچه مانند مدل‌های مبدل تصمیم می‌تواند راه‌های جدیدی را برای افزایش عملکرد با ترکیب یادگیری تقویتی و مدل‌سازی توالی ارائه دهد.

ایده ادغام یادگیری تقویتی با مدل‌سازی توالی در یک چارچوب یکپارچه، همانطور که در مبدل تصمیم‌گیری مشاهده می‌شود [21]، با اهداف تشخیص قسمت‌های برجسته ی اشیاء همسو می‌شود، جایی که وظیفه تمرکز پویا بر روی بخش‌های برجسته یک تصویر (یا ویدئو) برای تقسیم‌بندی و شی دقیق‌تر است. تشخیص مبدل‌های تصمیم، وظایف تصمیم‌گیری مانند تشخیص قسمت‌های برجسته ی اشیاء را به عنوان مشکلات پیش‌بینی توالی تلقی می‌کنند و رویکردی ساختاریافته‌تر و قابل تفسیر برای تشخیص مناطق برجسته را امکان‌پذیر می‌سازند. این به ویژه برای کاربردهایی مانند رانندگی مستقل یا دید رباتیک مفید است، جایی که شناسایی و تمرکز بر روی اشیاء برجسته در محیط برای تصمیم‌گیری ضروری است.

¹ Return to go

یکی از جنبه‌های کلیدی مدل‌های مبدل تصمیم [21]، مکانیسم بازگشت به رفتن است که به هدایت پیش‌بینی‌های عمل بر اساس پاداش‌های تجمعی کمک می‌کند. به طور مشابه، در تشخیص قسمت‌های برجسته ی اشیاء، یک شکل قابل مقایسه از بازگشت به رفتن می‌تواند برای هدایت تشخیص اشیاء برجسته بر اساس نتیجه مطلوب، مانند بهبود دقت یا افزایش کیفیت تقسیم بندی، استفاده شود. این شکل از شرطی سازی به مدل اجازه می‌دهد تا تمرکز خود را با ویژگی‌های خاص در یک تصویر بر اساس اهداف نهایی نقشه برجسته تطبیق دهد و برای طیف وسیعی از کاربردها، از گرفتن اشیا رباتیک تا ویرایش تصویر، انعطاف پذیری را ارائه می‌دهد.

مکانیسم‌های توجه مرکزی [20,22] برای مبدل‌های تصمیم نیز مشابهت‌های قوی در تشخیص قسمت‌های برجسته ی اشیاء پیدا می‌کنند. توجه به خود در مبدل‌های تصمیم‌گیری به مدل اجازه می‌دهد تا بر بخش‌های مهم یک دنباله، مانند اقدامات یا حالت‌های تصمیمات گذشته دور، تمرکز کند. در تشخیص قسمت‌های برجسته ی اشیاء، مکانیسم‌های توجه را می‌توان برای تمرکز بر ویژگی‌های مهم در یک تصویر به کار گرفت، که به طور بالقوه به مدل اجازه می‌دهد اشیاء برجسته را با توجه به نشانه‌های متنی محلی و جهانی شناسایی و بخش‌بندی کند. این می‌تواند به طور قابل توجهی دقت تشخیص برجسته را بهبود بخشد، به ویژه در صحنه‌های درهم و پیچیده.

علاوه بر این، مقیاس‌پذیری در مبدل‌های تصمیم [21]، استفاده از داده‌های مقیاس بزرگ و مجموعه داده‌های پیچیده را امکان‌پذیر می‌سازد، که برای بهبود استحکام مدل‌های تشخیص برجسته بسیار مهم است. با بزرگ‌نمایی مدل‌های تشخیص قسمت‌های برجسته ی اشیاء با استفاده از معماری‌های مبدل، می‌توان مجموعه‌های تصویری بزرگ و متنوع را مدیریت کرد و تعمیم مدل را در محیط‌های مختلف دنیای واقعی بهبود بخشید. در روشی مشابه، یادگیری آفلاین در مبدل‌های تصمیم [21] به مدل اجازه می‌دهد تا از داده‌های از پیش جمع‌آوری شده بدون نیاز به تعامل بلادرنگ یاد بگیرد. این مفهوم برای وظایف تشخیص قسمت‌های برجسته ی اشیاء در محیط‌هایی که بازخورد بلادرنگ امکان‌پذیر نیست، مانند تصویربرداری پزشکی یا تجزیه و تحلیل تصاویر ماهواره‌ای، که در آن آموزش بر روی مجموعه داده‌های گسترده‌ای از تصاویر مشروح می‌تواند منجر به مدل‌های تشخیص برجسته‌تر دقیق‌تر و کارآمدتر شود، ارزشمند است.

در نتیجه، ادغام تکنیک‌های مدل‌های مبدل تصمیم‌گیری [21] در تشخیص قسمت‌های برجسته ی اشیاء می‌تواند پیشرفت‌های چشمگیری را در هر دو زمینه ایجاد کند. با استفاده از چارچوب‌های یکپارچه،

مکانیسم‌های توجه و یادگیری آفلاین، تشخیص قسمت های برجسته ی اشیاء می‌تواند سازگارتر، مقیاس‌پذیرتر و کارآمدتر شود و کاربردهای خود را در محیط‌های پویا و دنیای واقعی گسترش دهد.

2-8- اصطلاحات مدل‌های مبدل تصمیم

اصطلاحات مرتبط با مدل‌های مبدل تصمیم [21] را می‌توان به حوزه تشخیص قسمت های برجسته ی اشیاء گسترش داد و درک دقیق‌تری از نحوه اعمال این رویکرد ترکیبی برای وظایف مبتنی بر تصویر و ویدیو را ممکن می‌سازد. اصول تصمیم‌گیری و مدل‌سازی توالی ذاتی در مبدل های تصمیم می‌تواند راحل‌های نوآورانه‌ای برای پیش‌بینی برجستگی ارائه دهد و توانایی مدل‌ها را برای شناسایی و بخش‌بندی مهم‌ترین اشیاء در یک صحنه افزایش دهد.

2-8-1- مسیر حرکت¹

در زمینه تشخیص قسمت های برجسته ی اشیاء ، یک مسیر به دنباله‌ای از فریم‌های تصویر یا بخش‌های ویدیویی اشاره دارد که هر کدام شامل حالات (ویژگی‌های بصری)، کنش‌ها (مناطق یا پیکسل‌های خاص تصویری که برای برجستگی هدف‌گذاری شده‌اند) و پاداش‌ها (اهمیت این مناطق بر اساس برجستگی) و دقت پیش‌بینی است. یک مسیر در تشخیص قسمت های برجسته ی اشیاء ممکن است مانند یک سری پیشنهادات شی شناسایی شده (وضعیت) به نظر برسد، که در آن اقدامات برجسته انجام می‌شود (پیش‌بینی مناطق برجسته)، و سیگنال های پاداش منعکس کننده دقت یا اثربخشی تشخیص برجسته هستند [21]. به عنوان مثال:

$$(S_1, a_1, r_1), (S_2, a_2, r_2), \dots$$

این مسیرها به عنوان ورودی برای آموزش مبدل های تصمیم برای پیش‌بینی مناطق برجسته با یادگیری از اقدامات و پاداش‌های فریم گذشته استفاده می‌شوند.

2-8-2- بازگشت به رفتن²:

در تشخیص قسمت های برجسته ی اشیاء ، بازگشت به رفتن نشان دهنده پاداش تجمعی تشخیص مناطق برجسته در فریم ها یا تصاویر آینده، مشروط به پیش‌بینی فعلی است. برای مثال، در مورد تشخیص

¹ Trajectory

² Return to go

برجستگی ویدیویی، بازگشت به رفتن می‌تواند مدل را برای تمرکز بر برجسته‌ترین شی در فریم‌ها، با در نظر گرفتن ویژگی‌های شی محلی و نشانه‌های متنی جهانی راهنمایی کند. این به مدل کمک می‌کند تا با یادگیری از تصمیمات برجسته گذشته و هدف نهایی تقسیم بندی شی، منطقه برجسته بعدی را پیش بینی کند. [21]

2-8-3- مدل سازی دنباله ای¹

تشخیص قسمت های برجسته ی اشیاء با استفاده از مبدل تصمیم می‌تواند به عنوان یک مشکل پیش بینی توالی در نظر گرفته شود، که در آن وظیفه پیش بینی توالی اشیاء یا مناطق برجسته در فریم ها یا تصاویر متعدد است. مدل یاد می‌گیرد که فریم‌های گذشته (حالت‌ها) و مناطق برجسته (اقدامات) شناسایی شده را به پیش‌بینی‌های آینده (مناطق برجسته) با در نظر گرفتن سیگنال بازگشت به رفتن، که مدل را به تمرکز بر مرتبط ترین بخش‌های صحنه هدایت می‌کند، نگاشت کند. این امر پیش‌بینی متوالی برجستگی را در طول زمان ممکن می‌سازد و دقت کلی تشخیص برجستگی پویا را بهبود می‌بخشد [21].

2-8-4- مکانیسم توجه²

مکانیسم توجه در مبدل های تصمیم نقش مهمی در تمرکز انتخابی بر روی بخش‌های کلیدی تصویر هنگام شناسایی اشیاء برجسته دارد. در تشخیص قسمت های برجسته ی اشیاء ، توجه به مدل اجازه می‌دهد تا به صورت پویا مهم ترین پیکسل ها یا مناطق را بر اساس زمینه صحنه و رابطه بین اشیاء مختلف برجسته کند. به عنوان مثال، این مدل می‌تواند به ویژگی‌های پس‌زمینه یا پیش‌زمینه تصویر توجه کند، و آن را قادر می‌سازد بر روی مناطقی با کنتراست بالا، مرزهای اشیا یا سایر نشانه‌های برجسته‌ای که برای تشخیص دقیق و تقسیم‌بندی اشیا حیاتی هستند، تمرکز کند [20,22].

2-8-5- تعبیه حالت³

نشان دهنده ویژگی های کلیدی یک تصویر یا فریم ویدیویی است که معمولاً از طریق شبکه های عصبی عمیق یاد می‌شود. این تعبیه‌ها ویژگی‌های معنایی و بصری یک صحنه، از جمله بافت، رنگ و مرزهای

¹ Sequence Modeling

² Attention Mechanism

³ State Embedding

شی را در بر می‌گیرد. با تبدیل داده‌های پیکسل خام به جاسازی‌های حالت، مبدل تصمیم می‌تواند این نمایش‌ها را به طور موثر پردازش کند و پیش‌بینی برجسته‌تر را امکان‌پذیر می‌سازد.[21]

2-8-6- اکشن جاسازی شده¹

یک عمل جاسازی شده در تشخیص قسمت های برجسته ی اشیاء نشان دهنده مناطق یا پیکسل های خاصی است که مدل به عنوان برجسته در یک تصویر یا فریم ویدئو مشخص می‌کند. با یادگیری جاسازی‌های این مناطق برجسته پیش‌بینی‌شده، مدل می‌تواند از معماری مبدل برای تمرکز بر حیاتی‌ترین عناصر بصری برای تشخیص برجسته‌تر استفاده کند. این تعبیه‌ها از طریق مکانیسم‌های توجه [20,22] و بازخورد پاداش اصلاح می‌شوند تا به تدریج فرآیند پیش‌بینی برجستگی را افزایش دهند [21].

2-8-7- سیاست شرطی پاداش²

سیاست شرطی پاداش مدلی است که شیء یا منطقه برجسته بعدی را بر اساس وضعیت فعلی (ویژگی های بصری تصویر) و پاداش تجمعی مورد انتظار (اهمیت یا ارتباط شی در صحنه) پیش‌بینی می‌کند. برخلاف مدل‌های تشخیص قسمت های برجسته ی اشیاء سنتی که بر ویژگی‌های دست ساز متکی هستند، سیاست شرطی پاداش در مبدل های تصمیم از سیگنال‌های بازگشت به تصمیم برای هدایت فرآیند تشخیص برجسته‌سازی استفاده می‌کند و تضمین می‌کند که مدل برجسته‌ترین اشیاء را بر اساس اهداف از پیش تعریف‌شده مانند دقت تقسیم‌بندی یا ارتباط بصری شناسایی می‌کند [21].

2-8-8- یادگیری تقویتی آفلاین³

یادگیری تقویتی آفلاین شامل مدل‌های آموزشی با استفاده از مجموعه داده‌های از پیش جمع‌آوری‌شده از تصاویر برجسته‌گذاری شده یا فریم‌های ویدئویی است که نیاز به تعامل بلادرنگ با محیط را از بین می‌برد. مبدل تصمیم می‌تواند از داده‌های تاریخی، مانند مجموعه داده‌های تصویر حاشیه‌نویسی در مقیاس بزرگ، برای یادگیری تشخیص برجسته بودن از صحنه‌های مختلف، بدون نیاز به بازخورد مداوم استفاده کند. این رویکرد به ویژه در کارهایی مانند تجزیه و تحلیل تصویر پزشکی یا دید خودروی خودمختار، که در آن جمع‌آوری داده‌های بلادرنگ می‌تواند پرهزینه یا غیرعملی باشد، ارزشمند است [21].

¹ Action Embedding

² Reward Conditioned Policy

³ Offline Reinforcement Learning

2-8-9- پنجره زمینه¹

پنجره زمینه به بخشی از مسیر (تصویر یا توالی ویدیو) اشاره دارد که مدل برای پیش‌بینی برجسته بودن در نظر می‌گیرد. در مورد تشخیص برجستگی، پنجره زمینه ممکن است چندین فریم را پوشش دهد، که به مدل اجازه می‌دهد هم از نشانه‌های بصری فوری (زمینه محلی) و هم از وابستگی‌های بلندمدت شی (زمینه جهانی) در سراسر دنباله یاد بگیرد. این مدل را قادر می‌سازد تا اشیاء برجسته‌ای را شناسایی کند که ممکن است تنها در طول زمان یا در زمینه اشیاء دیگر در صحنه برجسته شوند [21].

2-8-10- مبدل علی²

یک مبدل علی تضمین می‌کند که در کارهای پیش‌بینی برجستگی متوالی، مدل فقط از حالات گذشته و حال (تصاویر) برای پیش‌بینی آینده (اشیاء برجسته) استفاده می‌کند. در تشخیص قسمت های برجسته ی اشیاء ، این نوع مبدل را می توان برای ویدیوها یا دنباله هایی از تصاویر اعمال کرد و اطمینان حاصل کرد که پیش بینی های برجسته برای فریم های آینده تحت تأثیر اطلاعات آینده قرار نمی گیرند، و یک رابطه علی بین تصمیمات برجسته گذشته و پیش بینی های آینده حفظ می شود [21].

2-8-11- اکتشاف در مقابل بهره برداری³

در یادگیری تقویتی سنتی، مدل ها بین اکتشاف (آزمایش اقدامات جدید) و بهره برداری (با استفاده از اقدامات شناخته شده برای نتایج بهینه) تعادل برقرار می کنند. در تشخیص قسمت های برجسته ی اشیاء با استفاده از مبدل تصمیم، کاوش را می توان به عنوان جستجوی نشانه های برجسته جدید مشاهده کرد، در حالی که بهره برداری به تمرکز بر مناطق شناخته شده و برجسته در صحنه اشاره دارد. از آنجایی که مبدل های تصمیم به صورت آفلاین آموزش می بینند، بر تعادلی از کاوش مسیرهای مختلف (تصاویر) برای کشف ویژگی های برجسته قابل تعمیم تکیه می کنند در حالی که از آن ویژگی ها برای تشخیص دقیق برجستگی استفاده می کنند.

¹ Context Window

² Causal Transformer

³ Exploration vs. Exploitation

2-8-12- انتزاع زمانی¹

انتزاع زمانی به مبدل تصمیم اجازه می‌دهد تا وابستگی‌ها را در طول مراحل زمانی مختلف در مسیر (توالی فریم‌ها در ویدیوها) مدل‌سازی کند و مدل را قادر می‌سازد تا بر اساس زمینه بصری فوری و بلندمدت پیش‌بینی‌های برجسته انجام دهد. در وظایف تشخیص قسمت‌های برجسته ی اشیاء مانند تشخیص برجستگی ویدیویی، انتزاع زمانی به مدل کمک می‌کند تا چگونگی تکامل اشیاء در طول زمان را در نظر بگیرد و دقت تشخیص برجستگی را برای صحنه‌های پویا بهبود بخشد.

2-8-13- شبیه سازی سیاست²

شبیه‌سازی سیاست در تشخیص قسمت‌های برجسته ی اشیاء شامل آموزش مدلی برای تقلید از رفتار تشخیص برجسته بودن متخصص با یادگیری از مجموعه داده‌های برچسب‌گذاری شده با متخصص است. مبدل تصمیم می‌تواند از این مسیرهای متخصص بیاموزد و الگوهای پیش‌بینی برجستگی آن‌ها را تکرار کند، و اساساً فرآیند تولید نقشه برجسته‌سازی بهینه را با تقلید از تشخیص‌های برجستگی موفق ثبت‌شده در مجموعه داده تکرار می‌کند [21].

2-8-14- تنظیم دقیق³

تنظیم دقیق به مبدل‌های تصمیم‌گیری از پیش آموزش دیده اجازه می‌دهد تا با وظایف خاص تشخیص قسمت‌های برجسته ی اشیاء سازگار شوند. به عنوان مثال، یک مدل آموزش دیده در مورد وظایف برجسته عمومی را می‌توان با استفاده از یک مجموعه داده تخصصی (مثلاً تصاویر پزشکی یا فیلم‌های خودروی خودران) برای بهبود عملکرد در انواع خاصی از صحنه‌ها یا اشیاء تنظیم کرد. این باعث می‌شود که این مدل در برنامه‌های کاربردی دنیای واقعی موثرتر باشد و اطمینان حاصل شود که می‌تواند اشیاء برجسته را در زمینه‌های بسیار خاص تشخیص دهد.

¹ Temporal Abstraction

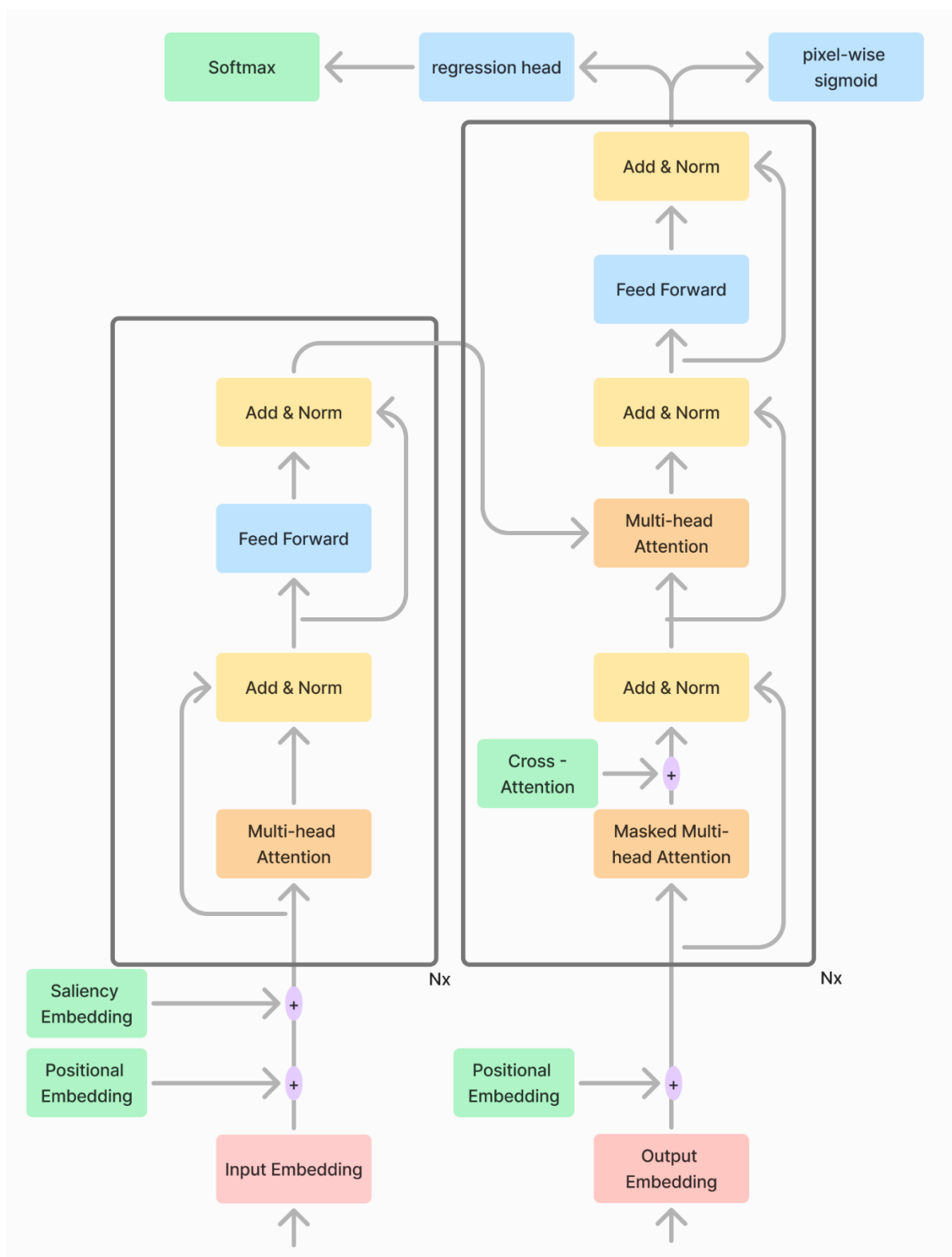
² Policy Cloning

³ Fine-Tuning

2-8-15- تعمیم¹

در نهایت، تعمیم، توانایی مدل برای انتقال دانش آموخته شده از یک مجموعه داده به سناریوهای جدید و دیده نشده است. در تشخیص قسمت های برجسته ی اشیاء ، این امر بسیار مهم است، زیرا مدل باید اشیاء برجسته را در طیف وسیعی از محیطها، شرایط نوری و انواع اشیاء شناسایی و بخش بندی کند. با استفاده از روش های آموزشی قوی مانند یادگیری تقویتی آفلاین و مکانیسم های توجه، مدل تصمیم گیری می تواند به موقعیت های جدید تعمیم داده و اشیاء برجسته را در انواع تنظیمات چالش برانگیز تشخیص دهد.

¹ Generalization



شکل (2-4) رویکرد تشخیص قسمت های برجسته اشیاء با مبدل تصمیّم

همانطور که در شکل 4-2 در معماری ترنسفورمر [20,21] برای تشخیص قسمت های برجسته ی اشیاء، ابتدا داده خام به بردارهای عددی (جاسازی ورودی¹) تبدیل شده و با اضافه شدن اطلاعات مکانی (جاسازی مکانی²) ترتیب فضایی داده ها حفظ می شود. برای تمرکز بر نواحی مهم تصویر، جاسازی قسمت های برجسته ی اشیاء³ اضافه می شود. مدل با استفاده از توجه چندسری⁴ روابط جهانی بین ویژگی ها را کشف کرده و از طریق جمع و نرمال سازی⁵ و لایه های تغذیه جلو⁶، ویژگی ها را بهبود می بخشد. این فرآیند در رمزگذاری⁷ چند بار تکرار می شود. در رمزگشایی⁸ نیز، پس از جاسازی خروجی و مکانی⁹، توجه چندسری ماسک شده ترتیب توالی را رعایت کرده و با توجه متقابل به خروجی رمزگذاری¹⁰، نواحی مرتبط با قسمت های برجسته ی اشیاء شناسایی می شوند. مراحل رمزگشایی مشابه رمزگذاری بوده و چندین بار تکرار می شود. در نهایت، نقشه قسمت های برجسته ی اشیاء¹¹ یا خروجی هایی مانند باکس ها¹² و برچسب های اشیاء¹³ تولید می شوند که هدف اصلی مدل را تشکیل می دهند. این معماری به مدل اجازه می دهد تا وابستگی های بلندمدت، نواحی مهم تصویر و روابط فضایی را به دقت شناسایی کند.

2-9- جمع بندی

در این فصل مفاهیم و تعاریف مورد استفاده در پژوهش معرفی و با جزئیات شرح داده شد که به بررسی ادبیات نظری و پیشینه تحقیق در زمینه تشخیص قسمت های برجسته ی اشیاء پرداخته شد و روش های سنتی شامل مبتنی بر کنتراست، پیشینه، و روش های هندسی، و روش های مبتنی بر یادگیری عمیق مانند شبکه های عصبی کاملاً کانولوشنی، شبکه های بازگشتی، و مکانیسم های توجه را مورد بررسی قرار داده شد. همچنین مدل های مبدل تصمیم که با استفاده از یادگیری تقویتی آفلاین و معماری های ترنسفورمر به

¹ Input Embedding

² Positional Embedding

³ Saliency Embedding

⁴ Multi-Head Attention

⁵ Add & Norm

⁶ Feed-Forward Layers

⁷ Encoder

⁸ Decoder

⁹ Output and Multi-Head Attention

¹⁰ Cross-Attention

¹¹ Map Saliency Object Detection

¹² Bounding Box

¹³ Object Label

شناسایی و بخش‌بندی دقیق اشیاء کمک می‌کنند، شرح داده شده است. مفاهیم کلیدی مانند مسیر حرکت، بازگشت به رفتن، مکانیسم‌های توجه، و تعمیم، به عنوان اصولی برای بهبود عملکرد در تشخیص برجستگی معرفی شده‌اند. در نهایت، به تحلیل و مقایسه روش‌های مختلف پرداخته شد. این فصل، پایه نظری جامعی برای فهم و توسعه روش جدید در فصول آتی فراهم می‌کند.

فصل 3: روش‌شناسی تحقیق

3-1- مقدمه

با توجه به اینکه در فصول قبل مفاهیم پایه و مطالعات پیشین بررسی شد، در این فصل نیز مدل مبدل تصمیم (Decision Transformer) جهت تشخیص قسمت های برجسته ی اشیاء مورد بررسی قرار می‌گیرد. اما قبل از اجرای مدل پیشنهادی، می‌بایست برخی پیش‌پردازش‌ها بر روی تصویر انجام شود تا نتایج با کارایی و سرعت بالا به دست آید. بنابراین در این فصل ابتدا به شرح مراحل پیش‌پردازش، سپس مدل و لایه مبدل تصمیم (ترنسفورمر) و در نهایت معماری مدل پیشنهادی (مبدل تصمیم) پرداخته می‌شود.

3-2- معماری مدل پیشنهادی

یک روش رایج برای پرداختن به تشخیص قسمت های برجسته ی اشیاء در بینایی کامپیوتر vision transformer (ViT) است که یک تصویر ورودی به مجموعه ای از وصله ها¹، نشانه‌هایی² از تصویر موردنظر که به بخش‌هایی مساوی تقسیم شده است [23,24]، تجزیه می‌کند و هر وصله از تصویر را در یک فضای بردار ترسیم می‌کند، و آن را با یک ضرب ماتریس به ابعاد کوچک تری نگاشت می‌کند به طوریکه وصله‌های مرتبط، نزدیک بهم باشند. سپس این تبعیه‌های برداری توسط رمزگذار ترنسفورمر طوری پردازش می‌شوند که گویی جاسازی رمزی هستند. در نهایت، یک مبدل برای تبدیل وصله‌ها به اصطلاح نشانه‌ها به فضای رمزگشا انتخاب می‌شود که به طور همزمان نقشه برجستگی [18] و نقشه مرزی [22] را از طریق نشانه‌های مربوط به وظیفه پیشنهادی و مکانیسم توجه وصله پیش‌بینی می‌کند. مشابه سایر روش‌های تشخیص قسمت های برجسته ی اشیاء مبتنی بر شبکه‌ها عصبی کانولوشنی، که اغلب از مدل‌های طبقه‌بندی تصویر از پیش آموزش دیده مانند VGG [7] و ResNet [8] به عنوان ستون فقرات³ رمزگذارهای خود برای استخراج ویژگی‌های تصویر استفاده می‌کنند، در اینجا هم مدل ViT⁴ [32] از پیش آموزش‌دیده را به عنوان ستون فقرات در نظر خواهیم گرفت.

مدل سازی ترنسفورمر تصمیم (مبدل تصمیم) [21] چشم انداز یادگیری تقویتی⁵ را با در نظر گرفتن یادگیری تقویتی به عنوان یک مسئله مدلسازی توالی شرطی [20,21] تغییر می‌دهد. مدل سازی ترنسفورمر

¹ patches

² tokens

³ backbone

⁴ Vision transformers

⁵ Reinforcement Learning

تصمیم به جای تکیه بر روش‌های سنتی یادگیری تقویتی، مانند برآزش یک تابع مقدار برای هدایت انتخاب کنش و به حداکثر رساندن بازده، از الگوریتم مدل‌سازی توالی (یعنی ترنسفورمر) برای تولید اقدامات آتی استفاده می‌کند که به بازده مورد نظر مشخصی دست می‌یابند. به عبارت ساده مبدل‌های تصمیم‌گیری نوعی مدل تخصصی ترنسفورمر را نشان می‌دهند که برای کارهایی که شامل تصمیم‌گیری گام به گام هستند ساخته شده است. این مدل‌ها در گرفتن توالی اطلاعات و تولید دنباله‌ای از اقدامات عالی هستند، که آنها را قادر می‌سازد تا تصمیمات آگاهانه را به شیوه‌ای ساختاریافته و متوالی اتخاذ کنند.[19,24]

این رویکرد نوآورانه به یک مدل خود رگرسیون مشروط به بازگشت مطلوب، وضعیت‌های گذشته و اقدامات وابسته است [20,22]. با استفاده از مدل‌سازی مسیر مولد، که الگوهای ترکیبی موقعیت‌ها، اقدامات و پاداش‌ها را پیش‌بینی می‌کند، این روش درک فرآیند یادگیری تقویتی را برای افراد ساده‌تر و آسان‌تر می‌کند، مبدل تصمیم‌گیری از فرآیند معمولی برای حداکثر کردن بازده دور می‌زند و مستقیماً یک مجموعه‌ای از اقدامات آینده که بازده مورد نظر را برآورده می‌کند.

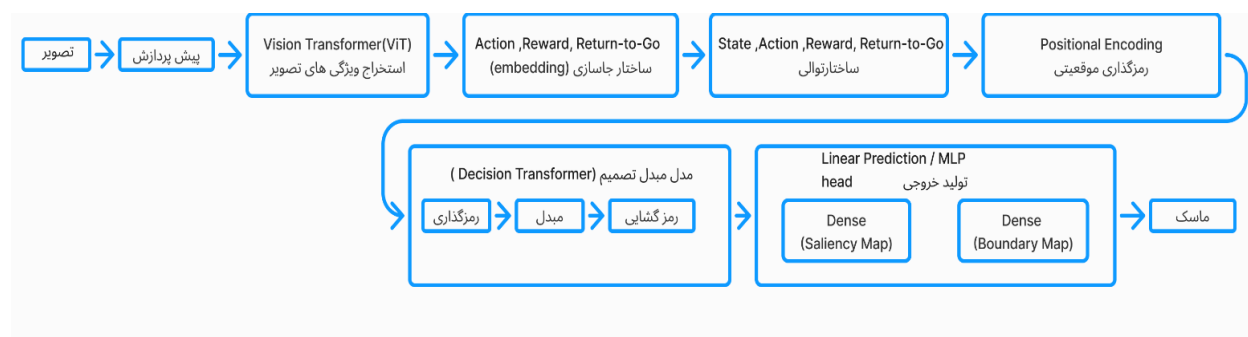
پس از استخراج ویژگی‌های تصویر با استفاده از ویژن ترنسفورمر [32] (ViT)، این ویژگی‌ها به همراه سایر اطلاعات مانند اقدامات، پاداش‌ها، بازگشت به آینده و زمانبندی‌ها¹ به مدل مبدل تصمیم داده می‌شوند. ابتدا، هر یک از این ورودی‌ها باید به فضای برداری مشابه تعبیه شوند. ویژگی‌های تصویر از طریق یک لایه خطی (Linear Layer) به یک فضای برداری با ابعاد مشخص مانند ۶۴ یا ۱۲۸ تعبیه می‌شوند. اقدامات که نشان‌دهنده خروجی هدف مدل هستند، از طریق یک لایه خطی دیگر به بردارهایی تبدیل می‌شوند که با سایر تعبیه‌ها قابل ترکیب باشند. پاداش‌ها که معیار کیفیت پیش‌بینی مدل در هر مرحله هستند، به صورت مقادیر عددی وارد می‌شوند و به کمک یک لایه خطی به فضای برداری تبدیل می‌شوند. بازگشت به آینده، که مجموع پاداش‌های آینده را نشان می‌دهد، نیز به همان روش تعبیه می‌شود. زمانبندی‌ها که موقعیت زمانی هر ورودی را نشان می‌دهند، از طریق یک لایه تعبیه‌سازی به بردارهایی با ابعاد مشابه تبدیل می‌شوند.

پس از انجام تعبیه‌سازی، تمامی این بردارها به صورت یک توالی ترکیب می‌شوند. به عنوان مثال، اگر طول توالی برابر با ۱۰ باشد و تعداد ویژگی‌های هر بردار ۶۴ باشد، نتیجه نهایی یک ماتریس با ابعاد [batch_size, sequence_length, hidden_size] خواهد بود. این توالی به بخش ترنسفورمر مدل داده

¹ TimeStep

می‌شود. در این مرحله، ترنسفورمر شامل چندین لایه رمزگذار است که هر لایه از دو بخش اصلی تشکیل شده است: بخش توجه چندسری¹ که ارتباط بین توکن‌های مختلف در توالی ورودی را مدل می‌کند، و بخش شبکه پیش‌خور² که ویژگی‌های تعبیه‌شده را با اعمال عملیات غیرخطی تقویت می‌کند. این فرآیند باعث می‌شود که اطلاعات وابستگی زمانی و محتوایی بین توکن‌ها بهتر درک شود.

در نهایت، خروجی ترنسفورمر که یک توالی از بردارها است، به یک لایه خروجی داده می‌شود. این لایه، خروجی‌هایی تولید می‌کند که تعداد آن‌ها با تعداد کلاس‌های اقدامات یا تعداد پیکسل‌های نقشه برجستگی برابر است. به عنوان مثال، اگر مدل برای پیش‌بینی نقشه برجستگی استفاده شود، خروجی نهایی یک ماتریس با ابعاد $[batch_size, height \times width]$ خواهد بود که مقادیر آن نشان‌دهنده احتمال برجستگی هر پیکسل است. این خروجی با استفاده از تابع هزینه‌ای مانند CrossEntropyLoss یا MAE با مقادیر واقعی مقایسه می‌شود و خطا محاسبه می‌گردد. گرادینان این خطا برای به‌روزرسانی وزن‌های مدل در مراحل بعدی استفاده می‌شود.



شکل (3-1) فرآیند کلی سیستم تشخیص قسمت‌های برجسته اشیاء

یکی از ویژگی‌های کلیدی مدل تصمیم‌گیری، ادغام اطلاعات چندسطحی از طریق توکن‌های رمزگذار است. ویژگی‌های سطح پایین، جزئیات دقیق را ارائه می‌دهند و ویژگی‌های سطح بالا اطلاعات زمینه‌ای و جهانی را فراهم می‌کنند. این دو نوع ویژگی در مرحله رمزگشایی ترکیب می‌شوند تا پیش‌بینی‌هایی هم دقیق و هم جامع ایجاد کنند. علاوه بر این، برای حفظ وضوح فضایی بالا در پیش‌بینی‌ها، این مدل از یک فرآیند نمونه‌برداری تدریجی استفاده می‌کند. در این روش، جاسازی‌های توکن به‌صورت تدریجی گسترش یافته و با وضوح تصویر ورودی اصلی هماهنگ می‌شوند تا مناطق برجسته به‌صورت دقیق و با جزئیات بالا شناسایی شوند. در نهایت، خروجی مدل شامل دو نقشه است: نقشه برجسته که با استفاده از یک سر

¹ Multi-Head Self-Attention

² Feedforward Network

رگرسیون و فعال‌سازی سیگموئید احتمال برجستگی پیکسل به پیکسل را پیش‌بینی می‌کند، و نقشه مرزی که با استفاده از سیگموئید زوجی روابط پیکسلی را تحلیل کرده و مرزهای دقیق اشیاء را شناسایی می‌کند. این دو خروجی به‌صورت مشترک آموزش داده می‌شوند تا فرآیند تشخیص مرزها به بهبود پیش‌بینی‌های برجسته‌سازی کمک کند.

ترنسفورمر تصمیم (DT) یک مدل توالی است که در اصل برای یادگیری تقویتی (RL) توسعه یافته است، که در آن مسیرها را به عنوان دنباله ای از حالت‌ها، اقدامات، پاداش‌ها، و بازگشت به رفتن (پاداش‌های آینده تجمعی) مدل می‌کند. برای گسترش این مدل برای ورودی‌های تصویر، ابتدا باید داده‌های تصویر در قالبی که با پردازش متوالی ترنسفورمر سازگار باشد، کدگذاری شود. تصاویر خام در هر مرحله زمانی معمولاً از یک ستون فقرات بصری مانند ResNet [8] یا Vision Transformer (ViT) [32] عبور می‌کنند تا تعبیه‌های ویژگی را استخراج کنند. این تعبیه‌ها حالت را در یک مسیر نشان می‌دهند. برای هر مرحله زمانی، یک تاپل مسیر متشکل از ویژگی‌های تصویر (وضعیت)، عمل، پاداش و بازگشت به رفتن ساخته، تعبیه شده و در یک دنباله الحاق می‌شود. رمزگذاری‌های موقعیتی برای حفظ نظم زمانی و آگاهی مکانی اضافه می‌شوند و اطمینان حاصل می‌کنند که ترنسفورمر می‌تواند بین مراحل زمانی و ترتیبات مکانی در داده‌های تصویر تمایز قائل شود.

سپس توالی اجزای تعبیه شده به ترنسفورمر وارد می‌شود، که با مدل‌سازی وابستگی‌های بین حالت‌های بصری و سایر عناصر مسیر، اقدام بعدی را پیش‌بینی می‌کند. این انطباق به DT اجازه می‌دهد تا وظایف RL مبتنی بر بینایی، مانند دستکاری رباتیک با ورودی‌های دوربین، و شبیه‌سازی رفتار از نمایش‌های متخصص شامل داده‌های بصری را حل کند. DT همچنین می‌تواند برای تصمیم‌گیری چند وجهی با ترکیب ورودی‌های تصویر با سایر روش‌ها مانند متن یا داده‌های عددی گسترش یابد. نقاط قوت آن شامل مقیاس پذیری برای مدیریت داده‌های بصری پیچیده، انعطاف پذیری در سیاست‌های آموزشی مشروط به سطوح مختلف پاداش، و یکپارچه‌سازی پردازش بصری و تصمیم‌گیری در یک معماری است. با این حال، با چالش‌هایی مانند هزینه‌های محاسباتی بالا، نیاز به مجموعه داده‌های بزرگ، و دشواری متعادل‌سازی نمایش‌های زمانی و مکانی در تصاویر مواجه است. با طراحی دقیق و بهینه‌سازی، DT پتانسیل قابل توجهی را برای اعمال نفوذ ورودی‌های تصویر در وظایف تصمیم‌گیری نشان می‌دهد.

3-2-1- مدل پیشنهادی

داده‌های مربوط به تشخیص قسمت‌های برجسته‌ی اشیاء که به صورت دو فایل جداگانه آموزش و تست و هر فایل به دو بخش تقسیم می‌شود: تصاویر و ماسک¹. داده‌ی ورودی پس از طی سه فاز مدل پیشنهادی، به خروجی مربوطه نگاشت می‌شود. فاز اول آنالیز داده‌ها، فاز دوم پیش‌پردازش تصاویر و ماسک، فاز سوم مدل Decision Transformer و فاز چهارم تولید خروجی می‌باشد. در ادامه به توضیح این 4 فاز پرداخته می‌شود.

3-2-1-1- آنالیز داده‌ها

هدف از تجزیه و تحلیل داده‌ها در این زمینه، درک ویژگی‌ها و کیفیت مجموعه داده، حصول اطمینان از مناسب بودن آن برای توسعه و ارزیابی مدل‌های تشخیص اشیاء برجسته است.

■ بارگذاری داده‌ها

ابتدا داده‌ها را بارگذاری می‌کنیم تا به محتویات آن دسترسی پیدا کنیم و سپس فرمت آن را بررسی می‌کنیم تا سازگاری و سازگاری را تأیید کنیم.

■ بررسی ابعاد داده‌ها

در مرحله بعد، ابعاد تصاویر و ماسک‌های مربوطه را بررسی می‌کنیم تا مطمئن شویم که آنها به درستی تراز هستند، زیرا ابعاد نامتناسب می‌تواند بر آموزش مدل تأثیر بگذارد. تجزیه و تحلیل مقادیر پیکسل در تصاویر و ماسک‌ها بینش‌هایی را در مورد توزیع داده‌ها ارائه می‌دهد و هرگونه ناهنجاری یا ناسازگاری را برجسته می‌کند. محاسبه ابعاد متوسط به شناسایی اندازه‌های معمولی کمک می‌کند، مراحل پیش‌پردازش مانند تغییر اندازه یا برش را راهنمایی می‌کند. محاسبه انحراف استاندارد پوشش برجسته در سراسر ماسک‌ها، حس شیوع‌شی در مجموعه داده را فراهم می‌کند و انتظارات مدل را مطلع می‌کند.

■ بررسی مقادیر پیکسل‌های منحصر به فرد

¹ mask

بررسی مقادیر پیکسل منحصر به فرد در ماسک‌ها تضمین می‌کند که برجسب‌ها به درستی رمزگذاری شده‌اند، که برای آموزش دقیق بسیار مهم است.

■ تشخیص نقاط پرت

در نهایت، تشخیص نقاط پرت به شناسایی نمونه‌های مشکل‌دار که می‌توانند نتایج را منحرف کنند، کمک می‌کند.

هر مرحله برای آماده‌سازی مؤثر مجموعه داده، اعتبارسنجی یکپارچگی آن، و تطبیق آن با الزامات وظایف تشخیص اشیاء برجسته ضروری است. برای اطمینان از کامل بودن، باید فایل‌های از دست رفته، حاشیه‌نویسی‌های ناقص و تعادل کلی مجموعه داده را نیز بررسی کنیم.

3-2-1-2- پیش‌آموزش

پس از انجام تجزیه و تحلیل داده‌ها و قبل از اعمال پردازش‌هایی از قبیل یافتن قسمت‌های برجسته اشیاء که پردازش‌های مقدماتی روی تصویر صورت گیرد، تا تفسیری قابل فهم برای ماشین تولید گردد. خروجی این مرحله نیز اهمیت زیادی برای افزایش کارایی سیستم و کاهش مرتبه زمانی دارد.

■ فرآیند تغییر اندازه

مرحله بعد از جمع‌آوری تصاویر¹، تغییر اندازه داده‌ها تصاویر و ماسک‌ها به ابعاد متوسط محاسبه‌شده در طول تجزیه و تحلیل تغییر اندازه داده می‌شوند تا اندازه‌های ورودی برای شبکه عصبی استاندارد شود، سربار محاسباتی کاهش می‌یابد و یکنواختی در کل مجموعه داده تضمین می‌شود.

■ نرمال‌سازی

مرحله بعد، نرمال‌سازی داده‌ها می‌باشد که مهم‌ترین عمل نرمال‌سازی تصویر استاندارد سازی پیکسل‌ها بین 0 و 1، که با اطمینان از اینکه ویژگی‌ها در مقیاس مشابه هستند، همگرایی مدل را بهبود می‌بخشد.

¹ Collecting Data

■ تقویت داده ها

گام بعد جهت افزایش مصنوعی تنوع داده‌ها و بهبود تعمیم مدل، تقویت داده‌ها می‌باشد که مهم‌ترین روش‌های تقویت داده تصویر عبارتند از:

- تبدیل های پیکسلی مانند: چرخش, برگرداندن, مقیاس بندی [43]
- فراتر از تبدیل‌های پیکسلی ساده تقویت داده در سطح نمونه با رنگ‌آمیزی مجدد بخش‌هایی از تصویر در سطح نمونه‌های شی است. [35]

■ مدیریت داده های نامتعادل

برای مجموعه داده‌های نامتعادل، تکنیک‌هایی مانند نمونه‌برداری بیش از حد از کلاس‌های اقلیت یا اعمال ضرر طبقات وزن‌دار برای رسیدگی به تفاوت‌ها در نمایش کلاس استفاده می‌شود.

■ تبدیل به تانسور

سپس، تمام داده‌ها به تانسور تبدیل می‌شوند، زیرا این فرمت برای مدل‌های PyTorch یا TensorFlow مورد نیاز است و پردازش و محاسبات کارآمد را در GPUها ممکن می‌سازد

■ نشانه گذاری

در نهایت، در این مرحله از پیش‌پردازش، متن به صورت عباراتی جدا از هم تقسیم می‌شود. هر عبارت، ممکن است فقط یک پیکسل یا چندین پیکسل که یک بخش از تصویر را تشکیل می‌دهند، باشد. می‌توان همه عبارات را تک پیکسلی و مستقل از هم و یا به صورت اصطلاحات چند پیکسلی در نظر گرفت.

داده‌ها طبق مراحل بیان شده، پیش‌پردازش می‌شوند، بدین صورت که هر مرحله پیش پردازش برای رسیدگی به مسائل خاص مجموعه داده، بهبود عملکرد مدل و آماده سازی داده ها برای یادگیری موثر ضروری است.

3-1-2-3- مدل مبدل تصمیم¹

مبدل تصمیم‌گیری (Decision Transformer) مفهوم بهینه‌سازی مسیر مبتنی بر یادگیری تقویتی را با تشخیص اشیاء برجسته تطبیق داده است. در این مدل، تولید نقشه برجستگی به‌عنوان دنباله‌ای از اقدامات تعریف می‌شود که هر یک از این اقدامات توسط سیگنال‌های پاداش هدایت می‌شوند. این مدل از معماری ترنسفورمر بهره می‌برد تا هم زمینه کلی تصویر و هم روابط محلی و فضایی را مدیریت کند. نتیجه این طراحی، مدلی است که توانایی تولید نقشه‌های برجستگی با کیفیت بالا و دقت قابل‌توجه را دارد.

فرآیند کار با تقسیم تصویر ورودی به تکه‌های غیرهمپوشان آغاز می‌شود. هر یک از این تکه‌ها به توکن‌هایی خطی تبدیل می‌شوند و با استفاده از کدگذاری موقعیتی، اطلاعات مکانی آن‌ها حفظ می‌شود. سپس، این توکن‌ها از طریق رمزگذار ترنسفورمر عبور داده می‌شوند که قادر است وابستگی‌های بلندمدت و روابط متنی میان توکن‌ها را استخراج کند و ویژگی‌هایی چندسطحی ایجاد نماید. در این ویژگی‌ها، توکن‌های سطح پایین جزئیات دقیق فضایی را حفظ می‌کنند، در حالی که توکن‌های سطح بالا اطلاعات معنایی و کلی تصویر را رمزگذاری می‌کنند. این ترکیب غنی از اطلاعات، نمایشی کامل از تصویر ورودی را برای مراحل بعدی فراهم می‌آورد.

در مرحله رمزگشایی، مبدل تصمیم‌گیری فرآیندی تدریجی و سلسله‌مراتبی برای اصلاح نقشه برجستگی ارائه می‌دهد. در این مرحله، نقشه برجستگی به‌عنوان دنباله‌ای از تصمیمات در نظر گرفته می‌شود که در هر گام، وضعیت فعلی نقشه برجستگی اصلاح و بهبود می‌یابد. وضعیت نقشه برجستگی با استفاده از توکن‌های حالت نمایش داده می‌شود که به طور پیوسته به‌روزرسانی می‌شوند. همچنین، توکن‌های اقدام طراحی شده‌اند تا تنظیمات لازم برای نقشه برجستگی را پیش‌بینی کنند. این پیش‌بینی‌ها از طریق تعامل میان توکن‌های حالت و ویژگی‌های رمزگذار، با استفاده از مکانیزم‌های توجه² انجام می‌شود. در این فرآیند، یک تابع پاداش بر اساس معیارهایی مانند IoU (تقاطع بر اتحاد) و دقت مرزی تعریف شده است. این پاداش‌ها مدل را هدایت می‌کنند تا در هر مرحله، نقشه‌ای دقیق‌تر و باکیفیت‌تر تولید کند. معماری کلی مدل ترنسفورمر بکار رفته در تشخیص قسمت های برجسته اشیاء در شکل (2-3) نمایش داده شده است.

¹ Decision Transformer² Attention Mechanism

3-2-1-4- تولید خروجی

در مدل مبدل تصمیم، فرآیند تولید خروجی مستقیماً از ویژگی‌های نهایی رمزگذار ترنسفورمر انجام می‌شود. این ویژگی‌ها که شامل اطلاعات معنایی و فضایی تصویر هستند، از طریق یک لایه خروجی خطی (nn.Linear) به فضای اقدام (Action Space) نگاشت می‌شوند. لایه خطی، ابعاد مخفی (Hidden Size) هر توکن را به تعداد اقدامات ممکن تبدیل می‌کند. خروجی نهایی مدل، مقادیر پیش‌بینی برای هر توکن است که بسته به مسئله مورد نظر، می‌تواند نشان‌دهنده دسته‌بندی پیکسل‌ها یا اقدام‌های مرتبط با نقشه برجستگی باشد.

در این معماری، برخلاف برخی مدل‌ها که از لایه‌های Sigmoid یا ماژول‌های Regression استفاده می‌کنند، مدل مبدل تصمیم خروجی‌ها را مستقیماً در فضای اقدام پیش‌بینی می‌کند. این رویکرد، پردازش ساده‌تر و متمرکزتری را برای مسائلی مانند طبقه‌بندی یا پیش‌بینی نقشه برجستگی ارائه می‌دهد. نتیجه خروجی در هر گام شامل پیش‌بینی مقادیر مرتبط با اقدامات ممکن برای هر توکن است و به مدل اجازه می‌دهد وابستگی‌های فضایی و معنایی تصویر را در نظر بگیرد.

3-2-2- آموزش مدل

برای آموزش مدل مبدل تصمیم، از بهینه‌ساز Adam با نرخ یادگیری¹ اولیه $1e-4$ استفاده شده است. همچنین، برای جلوگیری از بیش‌برازش و کنترل مقادیر پارامترها، مقدار $1e-5$ به‌عنوان وزن کاهشی² تنظیم شده است. در طول آموزش، تابع خطای CrossEntropyLoss به‌عنوان معیار اصلی بهینه‌سازی استفاده شده است. این تابع خطا خروجی‌های مدل را که در فضای اقدام قرار دارند، با مقادیر واقعی اقدامات مقایسه کرده و مقدار خطا را برای تنظیم وزن‌ها محاسبه می‌کند.

در فرآیند آموزش، تصاویر ورودی ابتدا از طریق ویژن ترنسفورمر (ViT) پردازش شده و ویژگی‌های مرتبط با آن‌ها استخراج می‌شود. این ویژگی‌ها به‌عنوان ورودی به مدل مبدل تصمیم داده می‌شوند. برای هر دسته از داده‌ها، اقدامات، پاداش‌ها، بازگشت به آینده و زمانبندی‌ها نیز به‌صورت جداگانه تعبیه شده و به ترنسفورمر ارسال می‌گردند. لایه‌های ترنسفورمر با استفاده از مکانیزم توجه چندسری و شبکه‌های

¹ Learning Rate

² Weight Decay

پیش‌خور، وابستگی‌های فضایی و معنایی میان ورودی‌ها را مدل کرده و بردارهایی نهایی تولید می‌کنند. این بردارها از طریق لایه خروجی به مقادیر پیش‌بینی در فضای اقدام نگاشت می‌شوند.

جهت ارزیابی عملکرد مدل در طول فرآیند آموزش، از معیار میانگین خطای مطلق¹ (MAE) برای بررسی کیفیت پیش‌بینی‌های مدل روی داده‌های اعتبارسنجی و آزمایشی استفاده شده است. آموزش با دسته‌های داده (Batch Size = 32) و برای ۵ دوره (Epoch) انجام شده است. مقادیر بهینه برای تنظیمات اولیه، از جمله نرخ یادگیری و تعداد گام‌ها، با استفاده از فرآیند آزمون و خطا تعیین شده‌اند. نتایج نشان داده‌اند که این مقادیر تنظیمی به تولید نقشه‌های برجستگی باکیفیت و پیش‌بینی دقیق کمک کرده‌اند. تمامی آزمایش‌ها و نتایج مرتبط با آموزش مدل در بخش بعدی و مقادیر پارامترها در جدول (3-1) گزارش شده است.

3-3- جمع‌بندی

در این فصل به شرح مدل پیشنهادی و مراحل ساخت مدل پرداخته شد. قبل از شرح مدل پیشنهادی، مراحل پیش‌پردازش متن به طور کامل توضیح داده شده و در ادامه به تفسیر لایه‌های استفاده شده در مدل پیشنهادی پرداخته شد و در نهایت معماری مدل پیشنهادی شرح داده شده است. فلوچارت مدل پیشنهادی و همچنین مقادیر پارامترها نیز در اواخر این فصل ارائه شده است. نتایج به‌دست آمده از اجرای این مدل نیز در فصل 4 ارائه شده است.

¹ Mean Absolute Error

پارامتر	مقادیر
نام مدل از پیش‌آموزش دیده	Vision Transformer
سایز بردار ورودی ¹ (ابعاد حالت)	768
ابعاد اکشن	50176
تعداد لایه خود توجه ²	2
تعداد سر های توجه ³	4
ابعاد فضای پنهانی	64
درصد احتمالی dropout	0.1
تابع خطا	MAE ⁴
نرخ یادگیری	4-1e
معیار ارزیابی مدل	MAE ⁵
بهینه‌ساز	AdamW
تعداد گام ⁶	10
اندازه دسته‌های ورودی ⁷	32
تعداد دوره‌ها ⁸	20
ابعاد خروجی مدل ⁹	متناسب با نقشه ی برجستگی

جدول (1-3) مقادیر پارامترها

¹ Input Dimension

² Self-Attention Layers

³ Attention Heads

⁴ Mean Absolute Error

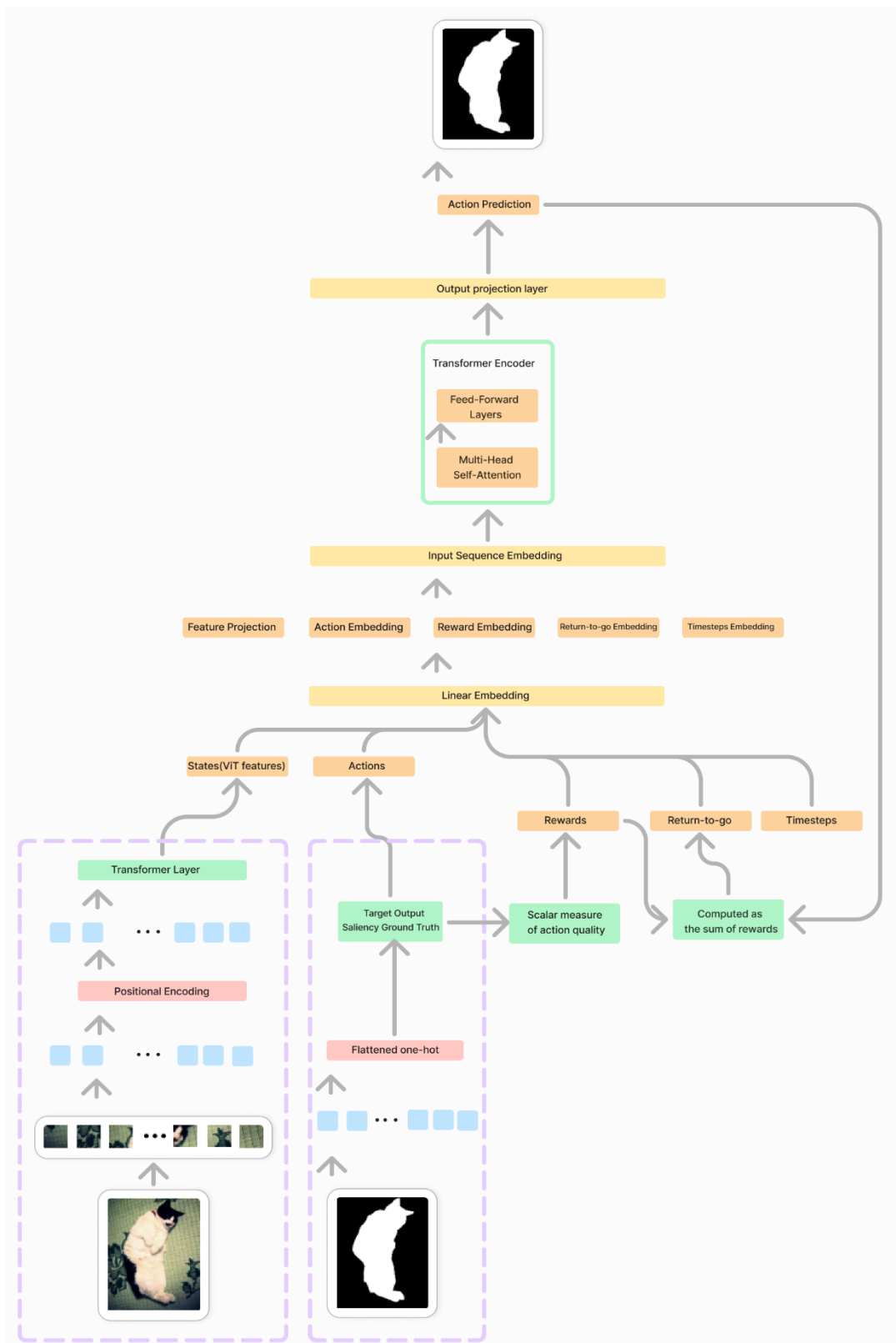
⁵ Mean Absolute Error

⁶ Sequence Length

⁷ Batch Size

⁸ Epochs

⁹ Action Dimension



شکل (2-3) معماری مدل پیشنهادی

فصل 4: یافته های تحقیق

4-1- مقدمه

در بخش های گذشته نحوه عملکرد روش Decision Transformer برای تشخیص قسمت های برجسته ی اشیاء معرفی شد. باتوجه به اینکه این روش، از مدل Transformer که در تسک های بینایی کامپیوتر از جمله کلاس بندی تصاویر¹، تشخیص شی² و تولید تصویر³ عملکرد خوبی دارد، استفاده کرده است. از عملکرد بالایی برخوردار می باشد. اما همچنان به دلیل محدودیت های موجود در تسک های بلادرنگ متوالی مثل رانندگی بدون سرنشین، مدیریت وظایف تصمیم گیری و ... به نتایج نزدیک به عملکرد بی وقفه انسان در شرایط تشخیص قسمت های خاص، دست نیافته است.

در این فصل نتایج حاصل از آزمایش عملی بر روی قسمت های برجسته اشیاء در تصویر، با استفاده از مدل پیشنهادی و دو روش موجود [32,33]، بررسی شده و نتایج بدست آمده از این آزمایشات مقایسه و ارزیابی می شود.

4-2- دادگان

با توجه به اینکه موضوع پژوهش، تشخیص قسمت های برجسته ی اشیاء می باشد، جمع آوری داده های مورد نیاز تصاویر مختلف از محیط اطراف می باشد. بدین منظور از دیتاست DUTS جهت دستیابی به داده های مورد نیاز استفاده شده است. DUTS یک مجموعه داده برای تشخیص قسمت های برجسته است که شامل 10553 تصویر آموزشی و 5019 تصویر آزمایشی است [43]. تمام تصاویر آموزشی از مجموعه های آموزشی ImageNet DET/val جمع آوری می شوند، در حالی که تصاویر آزمایشی از مجموعه آزمایش ImageNet DET و مجموعه داده SUN جمع آوری می شوند. هم مجموعه آموزشی و هم مجموعه تست شامل سناریوهای بسیار چالش برانگیزی برای تشخیص برجستگی است. اطلاعات حقیقی دقیق⁴ در سطح پیکسل به صورت دستی توسط 50 موضوع حاشیه نویسی می شود.

به دلیل کمبود داده در این زمینه، در قسمت پیش پردازش تعدادی داده نیز تولید می شود که به اصطلاح تقویت داده میگویند. تقویت داده ها معمولاً سوگیری های استقرایی در مورد فرآیند تشکیل تصویر را در آموزش گنجانده است (مانند ترجمه، مقیاس بندی، تغییر رنگ و...) که به روش تبدیل های پیکسلی در کل 10553 دیتا جدید برای آموزش و 5019 دیتا جدید برای حالت ارزیابی ایجاد کردیم. همچنین

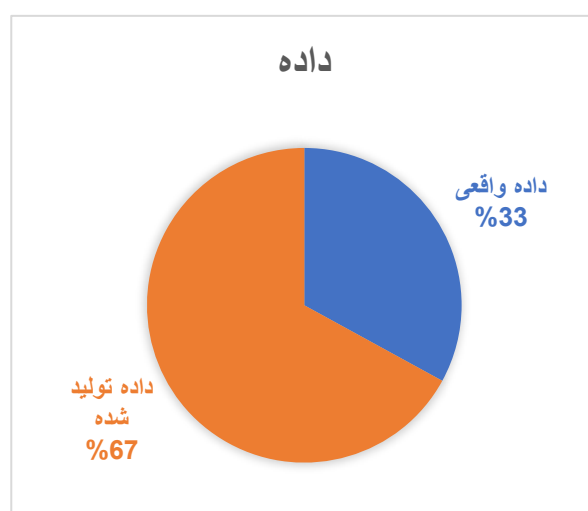
¹ Image Classification

² Object Detection

³ Image Generation

⁴ Ground Truth

میتوان از روش [35] برای تولید دیتا استفاده کرد که از مجموعه داده ی DUTS ، دیتای تقویت شده با 31656 تصویر و ماسک جدید را ایجاد میکند. این روش [35] یک مدل انتشار شرطی را با تهویه کنترل نقشه های عمق و لبه ترکیب می کند تا به طور یکپارچه اشیاء منفرد را در داخل صحنه رنگ آمیزی کند، که برای هر مجموعه داده های تقسیم بندی یا تشخیص ، قابل استفاده است. این روش [35] که به عنوان یک روش تقویت داده استفاده می شود، عملکرد و تعمیم مدل های تشخیص شی بر جسته، تقسیم بندی معنایی و مدل های تشخیص شی را بهبود می بخشد.





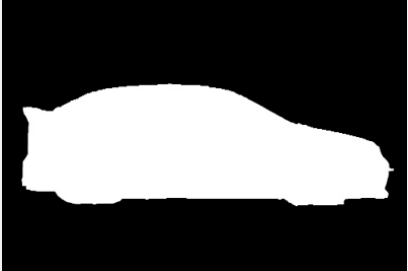





شکل (1-4) فراوانی داده ها

نام دسته	تعداد	درصد
آموزش	8442	55 %
ارزیابی	2111	13 %
تست	5019	32 %

جدول (1-4) تقسیم بندی داده ها برای مجموعه داده ی DUTS

نمونه ای از داده ها در جدول (1-4) و همچنین برای درک بهتر، فراوانی داده ها در شکل (1-4) موجود می باشد. همانطور که بیان شد ، دو نوع مجموعه داده وجود دارد ، مجموعه داده های واقعی می باشند از دیتاست [43] SUN , ImageNet ، و دومین مجموعه [35] ، شامل داده های تولید شده ، می باشد. جهت ارزیابی مدل دارای اهمیت فراوان می باشند. جهت آموزش مدل داده ها به سه دسته ی آموزش، ارزیابی و تست تقسیم شده اند. تعداد هر یک از این دسته ها در جدول (2-4) ارائه شده است.

نقشه برجستگی	تصویر	
		1
		2
	 <small>Copyright 2009 Menno Olgers</small>	3
		4

جدول (2-4) نمونه مجموعه داده ی DUTS

4-3- معیار های ارزیابی

به منظور ارزیابی موثر عملکرد روش پیشنهادی تشخیص قسمت های برجسته ی شی (SOD) با استفاده از مبدل های تصمیم، چندین معیار کلیدی در نظر گرفته شده است. برای ارزیابی اثربخشی سیستم های تشخیص قسمت های برجسته ی شی (SOD)، چندین معیار کمی استفاده می شود. این معیارها accuracy، precision و استحکام مدل ها¹ را در شرایط مختلف ارزیابی می کنند. که میتوان آنها را به چند گروه دسته بندی کرد.

4-3-1- معیارهای مبتنی بر منطقه²

معیارهای مبتنی بر منطقه در سیستم های تشخیص قسمت های برجسته شی (SOD) به صورت مستقیم کیفیت نقشه های برجسته پیش بینی شده را با نقشه های حقیقت زمین (Ground Truth) مقایسه می کنند. این ارزیابی در سطح پیکسلی انجام می شود، به این معنی که هر پیکسل در تصویر به صورت جداگانه بررسی می شود تا میزان انطباق پیش بینی مدل با واقعیت اندازه گیری شود.

4-3-1-1- دقت³ و یادآوری⁴:

دقت Precision نشان دهنده ی نسبت پیکسل های پیش بینی شده به عنوان برجسته که به درستی شناسایی شده اند که در رابطه 4-1 نشان داده شده است.

$$\frac{TP}{TP + FP} = p \quad (1-4)$$

که در آن TP تعداد پیکسل های مثبت صحیح و FP تعداد پیکسل های مثبت اشتباه هستند.

یادآوری Recall هم کسری از پیکسل های برجسته واقعی را که به درستی شناسایی شده اند را اندازه گیری می کند. در رابطه ی (4-2) فرموله شده است.

$$\frac{TP}{TP + FN} = R \quad (2-4)$$

¹ Robustness

² Region-based Metrics

³ Precision

⁴ Recall

که FN تعداد پیکسل های مثبت که شناسایی نشده اند را نشان می دهد. مقادیر هر کدام از جدول (4-1) بدست می آید:

مثبت	منفی	
منفی غلط پیش بینی شده (FP)	منفی درست پیش بینی شده (TN)	منفی
مثبت درست پیش بینی شده (TP)	مثبت غلط پیش بینی شده (FN)	مثبت

جدول (4-3) ماتریس درهم ریختگی

4-3-1-2 امتیاز F^1 :

یک میانگین هارمونیک وزنی از دقت و یادآوری است که در رابطه (4-3) فرمولاسیون این ارزیابی نشان داده شده است:

$$\frac{P \cdot R}{\beta^2 \cdot P + R} \cdot (\beta^2 + 1) = F \quad (4-3)$$

در این فرمول، $\beta > 1$ برای اولویت دادن به یادآوری و $\beta < 1$ برای اولویت دادن به دقت استفاده می شود. به طور معمول، برای اولویت دادن به دقت بر یادآوری استفاده می شود.

4-3-2-2 معیارهای مبتنی بر خطا

معیارهای مبتنی بر خطا برای ارزیابی میزان اختلاف و ناسازگاری بین نقشه برجسته پیش بینی شده و نقشه حقیقی (Ground Truth) استفاده می شوند. این معیارها به طور خاص برای اندازه گیری میزان خطا طراحی شده اند و به ما نشان می دهند که پیش بینی مدل چقدر از حالت ایده آل فاصله دارد.

4-3-2-1 میانگین خطای مطلق (MAE):

تفاوت پیکسلی بین نقشه برجسته پیش بینی شده و حقیقت زمین را ارزیابی می کند که در رابطه ی (4-4) تعریف می شود.

$$|G_i - P_i| \sum \frac{1}{N} = MAE \quad (4-4)$$

که در آن امتیاز برجستگی پیش‌بینی‌شده، امتیاز برجستگی حقیقت پایه و تعداد پیکسل‌ها است. P_i مقدار برجستگی پیش‌بینی‌شده برای پیکسل، G_i مقدار برجستگی در نقشه حقیقت زمین برای پیکسل و N تعداد کل پیکسل‌ها است.

4-3-3-3- معیارهای مبتنی بر ساختار

معیارهای مبتنی بر ساختار برای ارزیابی توانایی مدل در حفظ و شناسایی ساختارهای مکانی و روابط بین پیکسل‌ها در نقشه برجسته پیش‌بینی‌شده استفاده می‌شوند. برخلاف معیارهای مبتنی بر منطقه یا خطا که تمرکز اصلی‌شان روی کیفیت در مقیاس پیکسلی است، این معیارها به ارزیابی شباهت ساختاری و انسجام کلی بین نقشه پیش‌بینی‌شده و نقشه حقیقت زمین می‌پردازند.

4-3-3-1- اندازه‌گیری تشابه ساختاری (S_m):

شباهت ساختاری هر دو منطقه آگاه و شی آگاه بین نقشه برجسته پیش‌بینی شده و حقیقت زمین را ارزیابی می‌کند.

شباهت ساختاری آگاه به شی: تمرکز بر ساختار درونی اشیاء برجسته.

شباهت ساختاری آگاه از منطقه: مقایسه مناطق به عنوان یک کل.

ترکیب این دو شباهت به‌طور همزمان به ساختارهای کوچک (جزئیات) و مناطق بزرگ (کلیات) توجه دارد. که در رابطه (4-5) نشان داده شد.

(5-4)

$$S_r.(\alpha - 1) + S_o.\alpha = S_m$$

که در آن S_o شباهت ساختاری آگاه از شی، S_r شباهت ساختاری آگاه از منطقه و α یک ضریب وزنی که میزان اهمیت هر بخش را تنظیم می‌کند.

4-3-2-1- اندازه‌گیری تراز پیشرفته (E-measure):

خطاهای سطح پیکسل را با آمار سطح تصویر ترکیب می‌کند تا ارزیابی کند که نقشه برجسته پیش‌بینی‌شده چقدر با حقیقت زمین همسو می‌شود. این فرمول اطلاعات جهانی و محلی را ادغام می‌کند.

(6-4)

$$E(P_i, G_i) \Sigma \frac{1}{N} = E_m$$

که در آن $E(P_i, G_i)$ شاخص تراز بین مقدار برجستگی پیش‌بینی شده P_i و مقدار حقیقت زمین G_i است.

4-3-4- معیارهای مبتنی بر مرز

معیارهای مبتنی بر مرز برای ارزیابی توانایی مدل در شناسایی دقیق مرزهای اشیاء برجسته طراحی شده‌اند. این معیارها به‌جای تمرکز بر کل مناطق یا ساختارهای کلی، روی کیفیت تشخیص مرزهای اشیاء تمرکز دارند. مرزها معمولاً نواحی حساس و مهمی در تصاویر هستند، زیرا اطلاعات زیادی درباره شکل و موقعیت اشیاء ارائه می‌دهند.

4-3-4-1- خطای جابجایی مرزی (BDE^1):

میانگین جابجایی مرزهای پیش‌بینی شده را از مرزهای حقیقت زمین اندازه‌گیری می‌کند. مقادیر پایین‌تر نشان‌دهنده تراز بهتر است.

4-3-4-2- آگاه از مرز²:

بر روی اینکه مدل با چه دقتی مرزهای اشیاء برجسته را پیش‌بینی می‌کند، تمرکز می‌کند.

4-3-5- ارزیابی کلی

معیارهای ارزیابی در تشخیص قسمت‌های برجسته شی (SOD) به‌طور جامع عملکرد مدل را از جنبه‌های مختلف بررسی می‌کنند. معیارهای مبتنی بر منطقه مانند دقت، یادآوری، و (F-Measure) بر شناسایی صحیح و کامل مناطق برجسته تمرکز دارند، در حالی که معیارهای مبتنی بر خطا مانند میانگین خطای مطلق (MAE) که میزان اختلاف بین پیش‌بینی و حقیقت زمین را نشان می‌دهند. معیارهای مبتنی بر ساختار مانند S-measure و E-measure توانایی مدل در حفظ ساختارهای کلی و انسجام مکانی را می‌سنجند و معیارهای مبتنی بر مرز مانند خطای جابجایی مرزی یا BDE دقت در شناسایی مرزهای اشیاء را ارزیابی می‌کنند. ترکیب این معیارها یک چارچوب کامل برای بررسی جامع عملکرد مدل ارائه می‌دهد و امکان مقایسه دقیق بین مدل‌ها را فراهم می‌کند. اینجا ما معیارهای F-score، S-measure و MAE را جهت ارزیابی استفاده می‌کنیم.

¹ Boundary Displacement Error

² Boundary-Aware F-Measure

4-4- ارزیابی مدل پیشنهادی

در این بخش به ارزیابی مدل پیشنهادی پرداخته شده است. جهت ارزیابی مدل پیشنهادی دومدل دیگر [32,33] انجام گرفته است که در ادامه به توضیح هر آزمایش پرداخته می‌شود.

4-4-1- آزمایش اول

مدل Visual Saliency Transformer (VST) [32] شامل سه بخش اصلی است؛ یک رمزگذار ترنسفورمر، یک مبدل¹ و یک رمزگشای ترنسفورمر چندوظیفه‌ای. در رمزگذار از مدل پیش‌آموزش داده‌شده‌ی T2T-ViT² استفاده می‌شود که تصاویر را به بخش‌هایی تقسیم کرده و وابستگی‌های بلندمدت بین این بخش‌ها را پردازش می‌کند. مکانیزم T2T [32] در رمزگذار برای مدل‌سازی ساختارهای محلی و کاهش تدریجی طول توکن‌ها به کار می‌رود. مبدل، توکن‌های تولیدشده توسط رمزگذار را به فضای رمزگشا تبدیل می‌کند و در داده‌های RGB-D از Cross-Modality Transformer (CMT) برای ترکیب ویژگی‌های RGB و عمق استفاده می‌شود. در رمزگشا، مکانیزم جدید Reverse T2T (RT2T) برای افزایش تدریجی رزولوشن توکن‌ها به کار رفته و توکن‌های چندسطحی برای بهبود دقت پیش‌بینی‌ها ادغام می‌شوند. همچنین، توکن‌های مرتبط با وظایف، مانند برجستگی و مرزها، با استفاده از مکانیزم توجه³ تعامل دارند.

مدل از T2T-ViT پیش‌آموزش داده‌شده به عنوان بخش اصلی استفاده می‌کند که شامل ۱۴ لایه ترنسفورمر در رمزگذار و ابعاد جاسازی توکن‌ها (d) برابر با ۳۸۴ است. مبدل شامل ۴ لایه و رمزگشا شامل ۴ لایه در سطح و ۲ لایه در سطوح پایین‌تر است. اندازه دسته⁴ در آموزش برای داده‌های RGB برابر با ۱۱ و برای داده‌های RGB-D برابر با ۸ است. نرخ یادگیری اولیه مدل 0.0001 بوده که در طول آموزش کاهش می‌یابد و تعداد مراحل آموزش برای RGB و RGB-D به ترتیب ۴۰,۰۰۰ و ۶۰,۰۰۰ مرحله است. برای بهینه‌سازی از الگوریتم Adam استفاده شده است.

¹ Convertor

² Token-to-Token Vision Transformer

³ Patch-Task-Attention

⁴ Batch size

در فرآیند آموزش، از تابع خطای باینری کراس-انتروپی¹ برای پیش‌بینی برجستگی و مرزها استفاده شده است. نظارت عمیق در تمامی سطوح رمزگشا به کار گرفته شده و مکانیزم Patch-Task Attention بهبود آموزش را تسهیل کرده است.

در جدول زیر نتایج مدل VST جهت تشخیص قسمت های برجسته ی شی برای دیتاست DUTS ارائه شده است. معیارهای ارزیابی شامل معیار ساختاری (S_m^2)، امتیاز F بیشینه ($\max F$)، معیار بهینه‌سازی تقویت‌شده (E^{\max})، و خطای مطلق میانگین (MAE^3) هستن که در جدول (4-4) قرار داده شده است.

مدل	$S_m \uparrow$	$\max F \uparrow$	$E^{\max} \uparrow$	$MAE \downarrow$
VST	0.896	0.877	0.939	0.037

جدول (4-4) نتایج مدل VST

4-4-2- آزمایش دوم

این معماری [33] از ResNet-50 به عنوان شبکه پیش‌زمینه استفاده می‌کند که از طریق روش MoCo-v2 آموزش داده شده و نیازی به داده‌های برجسب‌دار دستی ندارد. ویژگی‌ها در بخش رمزگذار⁴ از لایه‌های مختلف استخراج شده و توسط بلوک‌های SE فشرده‌سازی و تحریک تلفیق می‌شوند. در بخش رمزگشا⁵ این ویژگی‌ها ترکیب شده و نقشه‌های برجستگی نهایی تولید می‌شوند. تصاویر ورودی به ابعاد 320×320 تغییر اندازه داده می‌شوند و خروجی معماری، نقشه‌های برجستگی است که نواحی مهم تصویر را نشان می‌دهد.

برای آموزش مدل، از نرخ یادگیری اولیه 0.1 که به صورت خطی کاهش می‌یابد، استفاده شده است. بهینه‌ساز به کار رفته گرادینان کاهشی تصادفی⁶ است و هر دسته داده شامل 8 نمونه است. فرایند آموزش برای 20 دوره اجرا می‌شود و شبکه‌های تشخیص اضافی نیز برای 10 دوره آموزش داده می‌شوند. در این مدل، سه تابع هزینه برای بهبود یادگیری تعریف شده است: تابع هزینه تفاوت حساسیت

¹ Binary Cross-Entropy

² S-Measure

³ Mean Absolute Error

⁴ Encoder

⁵ Decoder

⁶ Stochastic Gradient descent (SGD)

کنتر است¹ که به مدل کمک می کند به صورت تدریجی از نمونه های ساده به نمونه های دشوار یاد بگیرد، تابع هزینه پوشش زمانی پس از منیه² که مرزهای پیش بینی شده را با بافت های تصویر تطبیق می دهد و تابع هزینه سازگاری چندمقیاسی³ که پیش بینی های مدل را در مقیاس های مختلف هماهنگ می کند. وزن های این توابع به ترتیب برابر $\lambda_c=1$ ، $\lambda_b=0.05$ و $\lambda_m=1$ تنظیم شده و مقدار $\alpha=200$ برای فرآیند تطبیق بافت ها انتخاب شده است.

این مدل از استراتژی آموزشی پیش رونده استفاده می کند، به این معنا که یادگیری از نمونه های ساده آغاز شده و به نمونه های دشوار ختم می شود. همچنین، قابلیت پشتیبانی از داده های چندحالتی مانند تصاویر RGB، RGB-D، تصاویر حرارتی و ویدئو را دارد. برای بهبود کیفیت شبه برچسب ها (Pseudo-labels)، از روش میدان تصادفی شرطی⁴ به عنوان یک مرحله پس پردازش استفاده شده است.

عملکرد مدل با استفاده از سه معیار ارزیابی شده است، F_β (F-measure) که ترکیبی از دقت و بازخوانی است و بهترین مقدار آن $F_\beta=0.917$ بر روی مجموعه داده DUTS-TR ثبت شده است؛ میانگین خطای مطلق (MAE) که خطای پیکسلی را اندازه گیری کرده و مقدار بهینه آن برابر $M \approx 0.038$ است؛ و اندازه (E_ξ) که شباهت های محلی و جهانی را ارزیابی کرده و مقدار $E_\xi = 0.945$ به دست آمده است. در نهایت، با استفاده از ترکیب نهایی توابع هزینه (CSD + BTM + Multi-scale)، مدل بهبود عملکرد قابل توجهی را نشان داده و توانسته است نتایج برجسته ای با مقادیر $F_\beta=0.917$ ، $M=0.038$ و $E_\xi=0.945$ ارائه کند. این روش نشان می دهد که استراتژی های نوآورانه تعریف شده به شکل قابل توجهی توانایی مدل را در شناسایی و جداسازی اشیای برجسته بهبود بخشیده اند. در جدول (4-5) نتایج مربوط به مدل A2S-v2 درج شده است.

مدل	$E_\xi \uparrow$	$F_\beta \uparrow$	MAE \downarrow
A2S-v2	0.945	0.917	0.038

جدول (4-5) نتایج مدل A2S-v2

¹ Contrast Sensitivity Difference (CSD)

² Background Temporal Masking (BTM)

³ Local Mean Squared (LMS)

⁴ Conditional Random Field (CRF)

استراتژی آموزشی

مدل از استراتژی آموزشی پیش‌رونده استفاده می‌کند که در آن یادگیری از نمونه‌های ساده شروع شده و به نمونه‌های دشوار ختم می‌شود. همچنین، این روش قابلیت پشتیبانی از داده‌های چندحالتی مانند تصاویر RGB، RGB-D، تصاویر حرارتی و ویدئو را دارد. برای بهبود کیفیت برچسب‌های شبه (Pseudo-labels)، از روش CRF میدان تصادفی شرطی استفاده می‌شود.

4-4-3- آزمایش سوم (مدل مبدل تصمیم)

مدل Decision Transformer ترکیبی از ایده‌های مدل‌های ترنسفورمر و یادگیری تقویتی است که برای پیش‌بینی وضعیت‌ها، اقدامات و پاداش‌ها در یک چارچوب مبتنی بر توالی طراحی شده است. این مدل به‌طور خاص توالی‌هایی از داده‌ها به شکل $(R_1, S_1, a_1, R_2, S_2, a_2, \dots)$ را پردازش می‌کند و از ساختار ViT^1 برای استخراج ویژگی‌های تصویری استفاده می‌کند. ViT یا Vision Transformer یک مدل مبتنی بر ساختار ترنسفورمر است که به‌طور ویژه برای پردازش داده‌های تصویری طراحی شده است. برخلاف مدل‌های سنتی شبکه‌های عصبی کانولوشنی (CNNs) که برای پردازش تصاویر به‌کار می‌روند، ViT از ساختار ترنسفورمر، که معمولاً برای پردازش توالی‌ها استفاده می‌شود، برای پردازش تصاویر بهره می‌برد. مدل ViT ابتدا تصویر ورودی را به قطعات مربعی کوچک تقسیم می‌کند و سپس این قطعات به‌عنوان توالی ورودی به مدل داده می‌شوند. بعد از تبدیل این قطعات به بردارهای ویژگی، مدل ترنسفورمر روابط پیچیده میان بخش‌های مختلف تصویر را یاد می‌گیرد و ویژگی‌های تصویری را استخراج می‌کند. این ویژگی‌ها به‌عنوان ورودی به مدل داده می‌شوند تا بتوان توالی‌های مربوط به اشیاء برجسته را مدل‌سازی کرد.

ورودی‌های اصلی این مدل شامل حالت‌ها²، اقدامات³، بازده‌های آتی⁴ و زمان‌بندی‌ها⁵ هستند. حالت‌ها توسط شبکه Vision Transformer به بردارهایی با بعد ثابت تبدیل می‌شوند. اقدامات معمولاً به‌صورت مقادیر عددی یا بردارهایی با ابعاد مشخص ارائه می‌شوند و به فضای پنهانی بزرگ‌تر تعبیه می‌شوند. بازده‌های آتی نیز به‌عنوان سیگنالی برای هدایت مدل استفاده می‌شوند و در کنار زمان‌بندی‌ها به مدل

¹ Vision Transformer

² State

³ Action

⁴ Return-To-Go

⁵ TimeLine

ارائه می‌شوند تا ترتیب زمانی داده‌ها در نظر گرفته شود. برای پردازش این ورودی‌ها، هرکدام از آن‌ها از لایه‌های تعبیه استفاده می‌کنند تا با ابعاد فضای پنهانی مدل همخوان شوند.

مدل ترنسفورمر مورد استفاده در این پژوهش نسخه اصلاح‌شده‌ای از معماری Vision Transformer (ViT) است که به‌طور خاص برای پردازش داده‌های تصویری طراحی شده است. این مدل توانایی استخراج ویژگی‌های معنایی و فضایی از تصاویر ورودی را دارد و به‌عنوان بخشی از معماری مبدل تصمیم برای مدل‌سازی توالی‌هایی شامل بازده‌ها، حالت‌ها، و اقدامات مورد استفاده قرار می‌گیرد. برخلاف مدل‌های زبانی مانند GPT-2 که بر داده‌های متنی تمرکز دارند، این مدل با پیش‌آموزش بر داده‌های تصویری عظیم، توالی‌های بصری را مدل‌سازی می‌کند.

مدل شامل دو لایه ترنسفورمر رمزگذار با چهار head توجه در هر لایه است. ابعاد فضای پنهانی در این معماری 768 است و برای جلوگیری از بیش‌برازش، از نرخ افت تصادفی (Dropout) برابر با 0.1 در مکانیزم توجه و اتصالات باقی‌مانده استفاده می‌شود. همچنین، مدل شامل شبکه‌های پیش‌خور با ابعاد 3072 و فعال‌ساز GELU است که توانایی ترکیب اطلاعات زمانی و مکانی را به‌طور کارآمد فراهم می‌کند.

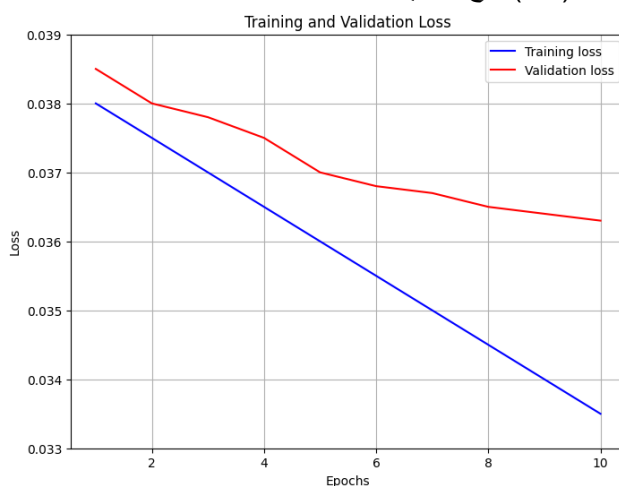
در این معماری، ورودی‌ها شامل ویژگی‌های استخراج‌شده از ViT، اقدامات، پاداش‌ها، بازده‌ها، و زمانبندی‌ها هستند که به‌صورت تعبیه‌شده در توالی‌های ورودی قرار می‌گیرند. ترنسفورمر با استفاده از مکانیزم توجه چندسری (Multi-Head Attention) و ترکیب اطلاعات زمانی و محتوایی، خروجی‌هایی به‌صورت پیش‌بینی اقدامات بعدی یا وضعیت‌های برجسته تصویر تولید می‌کند. این معماری نه‌تنها توانایی استخراج وابستگی‌های بلندمدت میان ورودی‌ها را دارد، بلکه امکان تولید نقشه‌های برجستگی دقیق و معنادار را نیز فراهم می‌آورد.

جدول (4-6) نتایج مدل پیشنهادی برای مقادیر مختلف پارامترها

ردیف	ابعاد اکشن	ابعاد حالت	تعداد لایه‌ها	تعداد head	ابعاد فضای پنهانی	سایز بسته	تعداد گام	نرخ یادگیری	Dropout	MAE	F-measure
1	50176	768	2	4	64	32	10	10^{-5}	0.1	0.036	0.95
2	50176	768	4	8	128	32	10	10^{-5}	0.1	0.035	0.95

برای ارزیابی عملکرد مدل، معیارهای مختلفی بسته به نوع خروجی‌ها به کار گرفته می‌شوند. برای پیش‌بینی وضعیت‌ها، که در اینجا به صورت نقشه‌های برجستگی (saliency maps) است، معیارهایی مانند میانگین خطای مطلق (MAE) و شاخص معیار اف (F_β) استفاده می‌شوند. این معیارها دقت مدل در شناسایی بخش‌های برجسته تصویر را می‌سنجند. برای پیش‌بینی اقدامات و بازده‌ها، معمولاً معیارهای خطای میانگین مربعات (MSE) یا خطای مطلق (MAE) به کار می‌روند.

شکل (2-4) نتایج مدل پیشنهادی روی داده های آموزش و ارزیابی



مدل Decision Transformer با ترکیب قدرت یادگیری تقویتی و مدل‌های ترنسفورمر، قابلیت انعطاف‌پذیر برای یادگیری توالی‌های مختلف ارائه می‌دهد. این مدل نیازی به تعریف مستقیم توابع ارزش یا سیاست ندارد و مستقیماً از داده‌ها یاد می‌گیرد. یکی از کاربردهای اصلی این مدل، وظایف مرتبط با یادگیری تقویتی در محیط‌های پیچیده است، اما انعطاف آن امکان استفاده در مسائل بینایی، مانند شناسایی اشیاء برجسته، را نیز فراهم می‌کند. چالش اصلی این مدل، هزینه محاسباتی بالا و نیاز به آماده‌سازی دقیق داده‌ها است. با این حال، توانایی مدل در یادگیری مستقیم از داده‌ها آن را به ابزاری قدرتمند برای کاربردهای متنوع تبدیل کرده است.

جدول (7-4) نتایج T-Test آزمایش سوم

مدل	MAE	F-Measure
<i>DT</i>	0.036	0.94

4-5- مقایسه روش مبدل تصمیم با روش های موجود

مدل مبدل تصمیم که در این پروژه استفاده شده است، رویکردی نوآورانه برای شبیه سازی توالی های بازده، وضعیت و اقدام به طور خودبازگشتی است. در مقایسه با روش های موجود در مقالات [33] و [32]، مدل Decision Transformer یک مزیت کلیدی دارد. این مزیت در پردازش توالی های طولانی از داده ها و یادگیری مستقیم از توالی های زمان بندی شده نهفته است، که نیازی به تعریف مستقیم توابع ارزش یا سیاست ندارد.

در حالی که مدل [33] برای کشف ویژگی های بافتی در تصاویر و استفاده از آن ها برای شناسایی اشیای برجسته طراحی شده است، و مدل [32] از قدرت ترنسفورمر برای شبیه سازی توجه به اشیای برجسته در تصاویر استفاده می کند، مدل مبدل تصمیم (Decision Transformer) قادر است تا با ترکیب ویژگی های مختلف از جمله ویژگی های تصویری و بازده ها، یک مدل خودآموز و تطبیقی برای پیش بینی اقدامات و وضعیت ها ایجاد کند.

یکی از نقاط ضعف مدل های موجود مانند [32] این است که بر اساس پردازش تصاویر به صورت مجزا عمل می کنند و ممکن است نتوانند به طور موثر با داده های با ابعاد بزرگ یا داده های پیچیده تر از جمله توالی های زمانی که در تصمیم گیری های پیچیده دخیل هستند، تعامل کنند. مدل Decision Transformer از آن جا که یادگیری از توالی های چند بعدی را تسهیل می کند، قادر است به طور موثر در محیط های پیچیده با استفاده از ویژگی های زمانی و تصویری به بهترین نتایج برسد.

4-6- تحلیل نتایج

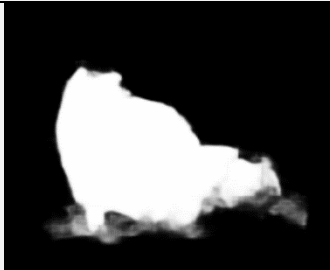


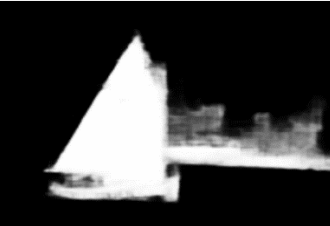
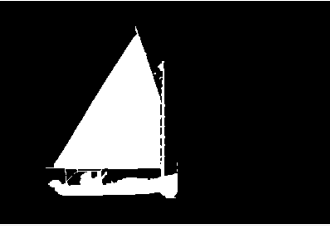







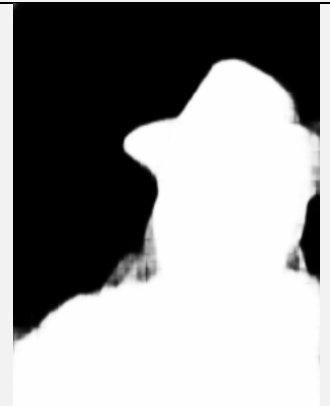
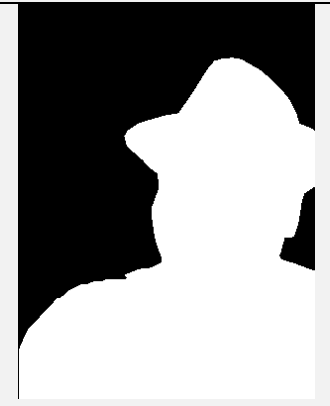

نتایج مدل مبدل تصمیم (Decision Transformer) نشان می دهد که این مدل توانایی بالایی در شبیه سازی و پیش بینی وضعیت ها، اقدامات و بازده ها در یک چارچوب یادگیری تقویتی دارد. این مدل به ویژه در پردازش داده هایی که شامل تعاملات زمانی پیچیده و ویژگی های تصویری هستند، عملکرد چشمگیری از خود نشان داده است. در مقایسه با سایر مدل ها، عملکرد آن در معیارهایی مانند MSE، S-measure و F-measure قابل توجه است. این انتخاب از معیارها به ویژه در سنجش دقت پیش بینی های مدل و ارزیابی توانایی آن در شبیه سازی تعاملات زمانی و ویژگی های تصویری مناسب است. MSE (Mean Squared Error). برای ارزیابی تفاوت بین مقادیر پیش بینی شده و مقادیر واقعی در بازده ها و اقدامات استفاده می شود، در حالی که S-measure و F-measure به طور خاص برای

سنجش دقت و کیفیت شبیه‌سازی ویژگی‌های تصویری مانند دقت اشیای برجسته (salient objects) به‌کار می‌روند.

مدل مبدل تصمیم با استفاده از معماری ترنسفورمر (Transformer) و توانایی مدل‌سازی وابستگی‌های زمانی و مکانی، به‌طور مؤثری توانسته است از ویژگی‌های مختلف به‌ویژه ویژگی‌های تصویری و زمانی بهره‌برداری کند. انتخاب این معماری به دلیل قدرت بالای آن در پردازش توالی‌ها و استفاده از لایه‌های توجه (Attention Layers) در کنار قابلیت‌های خاص آن در زمینه یادگیری غیرمستقیم و مدل‌سازی روابط پیچیده بین داده‌ها صورت گرفته است. یکی از ویژگی‌های مهم این مدل تعداد نرون‌ها و پارامترهای بهینه‌شده است. به‌عنوان مثال، مدل Decision Transformer با تعداد ۱۲ لایه و ۱۲ head توجه برای هر لایه طراحی شده است که به‌طور مؤثری توانسته است هم‌زمان از تعاملات پیچیده زمانی و ویژگی‌های تصویری استفاده کند. انتخاب این تعداد نرون و لایه‌ها به‌منظور تعادل میان دقت پیش‌بینی و جلوگیری از پیچیدگی بیش‌ازحد و زمان آموزش طولانی بوده است. استفاده از تعداد نرون‌های معین باعث می‌شود که مدل قادر به پردازش و یادگیری ویژگی‌های پیچیده در داده‌ها باشد، بدون اینکه از نظر محاسباتی بهینه‌سازی مدل تحت تاثیر قرار گیرد.

علاوه بر این، بهینه‌سازی مدل از طریق الگوریتم Adam (که یک بهینه‌ساز مبتنی بر گرادیان است) صورت گرفته است. انتخاب این بهینه‌ساز به‌دلیل ویژگی‌های آن در یادگیری سریع و دقیق پارامترهای مدل بوده است. Adam با به‌روزرسانی نرخ یادگیری به‌طور خودکار و با استفاده از تاریخچه مقادیر گرادیان، قادر است به‌طور مؤثری بهینه‌سازی مدل را در زمینه‌های مختلف یادگیری ماشین انجام دهد. این انتخاب باعث افزایش پایداری و سرعت آموزش مدل شده است، به‌ویژه در شرایطی که داده‌های پیچیده و متنوع با حجم زیاد در دسترس هستند.

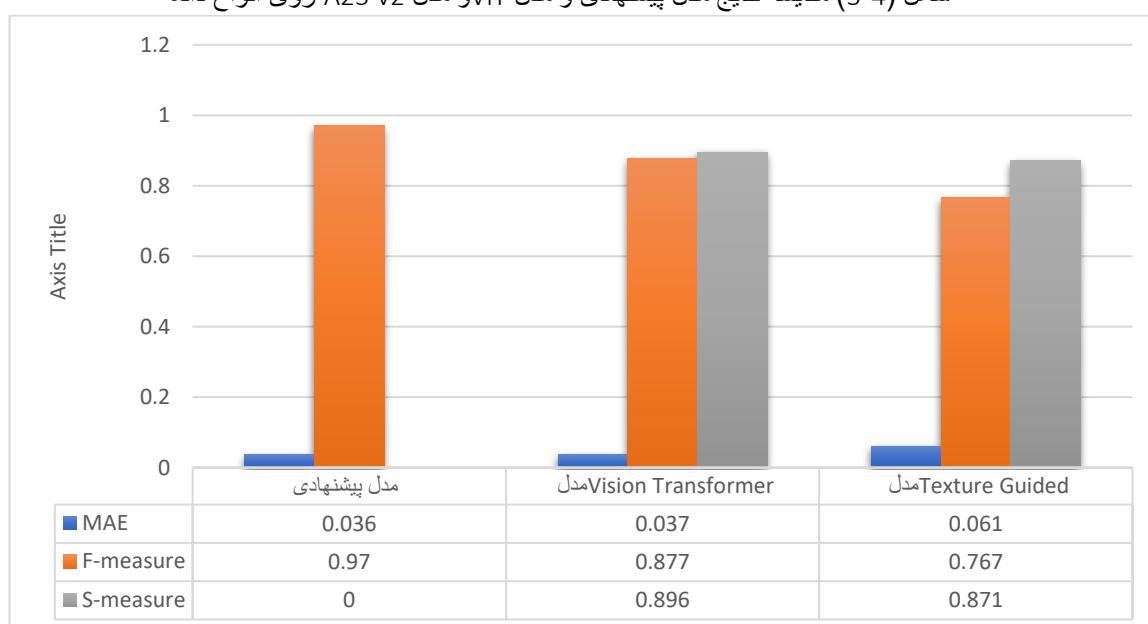
در مقایسه با مدل [32] که از لایه‌های خود توجه (Self-attention Layers) برای شبیه‌سازی توجه به ویژگی‌های برجسته استفاده می‌کند، مبدل تصمیم (Decision Transformer) قادر است تا از ورودی‌های چندگانه (شامل ویژگی‌های تصویری و زمانی) برای مدل‌سازی توجه به اشیای برجسته به‌طور خودکار استفاده کند. این ویژگی به‌ویژه در شرایطی که تعاملات زمانی پیچیده و داده‌های تصویری متنوع وجود داشته باشد، از اهمیت ویژه‌ای برخوردار است. مدل Decision Transformer با استفاده از داده‌های زمان-محور به‌طور مؤثری می‌تواند وابستگی‌های طولانی‌مدت را شبیه‌سازی کرده و پیش‌بینی‌هایی دقیق‌تر ارائه دهد. که در جدول (4-9) 5 مورد از پیش‌بینی‌های خروجی مدل پیشنهادی ما را مشاهده می‌کنید.

نقشه برجستگی - پیش بینی شده	نقشه برجستگی - واقعی	تصویر	
			1
			2
			3
			4
			5

جدول (8-4) نمونه مجموعه داده ی DUTS

در نهایت، مدل مبدل تصمیم با استفاده از ویژگی‌های متنوع و پیکربندی انعطاف‌پذیر توانسته است عملکرد مناسبی در مقایسه با دیگر مدل‌ها از جمله مدل‌های بافتی¹ [33] و ترنسفورمری² [32] نشان دهد. این توانمندی در بهره‌برداری از داده‌های تصویری و ویژگی‌های زمانی به‌طور هم‌زمان، باعث برتری آن در شبیه‌سازی اشیای برجسته در تصاویر و پیش‌بینی‌های دقیق‌تر شده است. انتخاب این ترکیب از ویژگی‌ها و پارامترها به‌طور خاص به دلیل آن است که مدل قادر به استفاده از داده‌ها و ویژگی‌های مختلف برای مدل‌سازی بهتر فرآیندهای پیچیده در شبیه‌سازی و پیش‌بینی رفتارهای اشیای برجسته در تصاویر است.

شکل (3-4) مقایسه نتایج مدل پیشنهادی و مدل ViT و مدل A2S-v2 روی انواع داده



7-4- جمع‌بندی

در این تحقیق، مدل Decision Transformer به‌عنوان یک مدل پیشرفته برای شبیه‌سازی توالی‌های پیچیده و پیش‌بینی وضعیت‌ها، اقدامات و بازده‌ها در محیط‌های مبتنی بر یادگیری تقویتی معرفی شد. این مدل با استفاده از ویژگی‌های تصویری استخراج‌شده از ViT و بازده‌های آتی، به‌طور خودکار توالی‌های زمانی را پردازش کرده و تصمیمات مناسبی اتخاذ می‌کند. مقایسه مدل Decision Transformer با مدل‌های موجود مانند [33] A2S-v2 و [32] ViT نشان می‌دهد که این مدل قادر است در شناسایی اشیای برجسته با دقت بالا و کارایی بهتر در محیط‌های پیچیده عمل کند. نتایج نشان‌دهنده

¹ Texture-guided

² Visual Saliency

عملکرد مناسب این مدل در معیارهای مختلف است، که موجب پیشنهاد آن به عنوان یک روش مؤثر در شبیه سازی و پیش بینی وضعیت ها در مسائل شناسایی اشیا برجسته می شود.

با توجه به این که مدل Decision Transformer توانسته است توانمندی های بالایی در مقایسه با روش های سنتی ارائه دهد، این تحقیق به عنوان یک گام مهم در استفاده از مدل های ترنسفورمر برای مسائل پیچیده تر شناسایی اشیا برجسته و یادگیری تقویتی در نظر گرفته می شود.

فصل 5: نتیجه‌گیری و ارائه پیشنهاد

5-1- نتیجه‌گیری

این تحقیق بر روی تشخیص بخش‌های برجسته اشیاء درون تصاویر تمرکز دارد که یک مشکل اساسی و دیرینه در بینایی کامپیوتر و هوش مصنوعی است. هدف این کار با تقلید از سیستم‌های بصری انسان، این است که رایانه‌ها را قادر به شناسایی، تجزیه و تحلیل و اولویت‌بندی مناطق مهم بصری در تصاویر کند. این مطالعه از مبدل‌های تصمیم (DT)، یک مدل جدید و نوآورانه با الهام از یادگیری تقویتی، برای رسیدگی به چالش‌های موجود در تشخیص شیء برجسته (SOD) استفاده می‌کند. برخلاف روش‌های سنتی، که به شدت بر ویژگی‌های از پیش تعریف‌شده و قوانین ثابت تکیه می‌کنند، رویکرد DT از تجربیات آموخته‌شده، سازگاری و مکانیسم‌های مبتنی بر پاداش برای بهبود دقت و کارایی بدون بازخورد ثابت استفاده می‌کند.

این تحقیق بر کاربرد DT در سناریوهای بلادرنگ و مقیاس بزرگ با هدف متعادل کردن کارایی محاسباتی با دقت تاکید دارد. این ادغام اطلاعات زمینه‌ای، مانند روابط فضایی، انسجام معنایی، و مکانیسم‌های توجه را بررسی می‌کند تا قابلیت‌های تشخیص را افزایش دهد. با معرفی فرآیندهای یادگیری مبتنی بر پاداش، DT ها انتشار خطا را کاهش می‌دهند و استحکام را نشان می‌دهند و راه را برای مدل‌های قابل اعتماد تشخیص اشیاء مناسب برای محیط‌های متنوع هموار می‌کنند.

این مطالعه یک تجزیه و تحلیل مقایسه‌ای کامل از DT ها را در برابر تکنیک‌های یادگیری عمیق معمولی و مدرن ارائه می‌دهد و سازگاری، استحکام و دقت برتر آنها را در سناریوهای پیچیده و پویا نشان می‌دهد. کاربردهای عملی در زمینه‌هایی مانند وسایل نقلیه خودران نشان داده شده است، جایی که تشخیص دقیق و کارآمد اشیاء حیاتی ضروری است. تصویربرداری پزشکی، برای بهبود تشخیص و کاهش خطا؛ و سیستم‌های تعاملی، افزایش تعامل انسان و کامپیوتر. این کاربردها بر اهمیت این تحقیق به عنوان گامی دگرگون‌کننده در پیشرفت فناوری‌های بینایی کامپیوتر تاکید می‌کند.

5-2- کارهای آینده

■ بهینه‌سازی مدل پیشرفته

تحقیقات آینده می‌تواند بر کاهش سربار محاسباتی مبدل‌های تصمیم تمرکز کند و آنها را برای استقرار در دستگاه‌های دارای محدودیت منابع مانند تلفن‌های همراه، دستگاه‌های لبه و سیستم‌های تعبیه شده مناسب

سازد. کاوش در معماری مبدل های سبک وزن، تکنیک های کوانتیزاسیون¹ یا هرس مدل می تواند به طور قابل توجهی کارایی آنها را در عین حفظ عملکرد افزایش دهد.

■ ادغام با داده های چندوجهی

گسترش چارچوب فعلی برای ترکیب داده های چندوجهی (به عنوان مثال، ترکیب اطلاعات بصری، شنیداری، متنی و حسی) می تواند توانایی مدل را برای درک و تفسیر صحنه های پیچیده، به ویژه در محیط های دنیای واقعی که انواع داده های متنوع در کنار هم وجود دارند، بهبود بخشد.

■ یادگیری متنی پیشرفته

بررسی مکانیسم های پیشرفته برای ادغام آگاهی زمینه ای عمیق تر، مانند وابستگی های زمانی و مکانی دوربرد در داده های ویدئویی، می تواند عملکرد مدل را در محیط های پویا و در حال تکامل بیشتر بهبود بخشد. استفاده از مکانیسم های توجه مکانی-زمانی می تواند یک جهت امیدوارکننده باشد.

■ استقرار و آزمایش در دنیای واقعی

استقرار و آزمایش دقیق این مدل در کاربردهای دنیای واقعی، مانند ناوبری خودکار، تشخیص پزشکی، رباتیک و واقعیت افزوده می تواند عملی بودن آن را تأیید کند. این آزمون ها به شناسایی محدودیت های خاص کمک می کنند و بینش هایی را برای اصلاحات تکراری ارائه می کنند.

■ پرداختن به ذهنیت در برجستگی

ادراک برجستگی به دلیل عوامل فرهنگی، روان شناختی و زمینه ای می تواند در بین افراد بسیار متفاوت باشد. کار آینده می تواند مدل های تشخیص برجسته شخصی یا تطبیقی را که ترجیحات خاص کاربر را از طریق بازخورد تعاملی یا مکانیسم های تطبیقی یاد می گیرند، کشف کند.

■ تعمیم بهبود یافته

بررسی راه هایی برای افزایش قابلیت های تعمیم مدل به مجموعه داده های دیده نشده و سناریوهای جدید، آن را در بین برنامه های مختلف دنیای واقعی متنوع تر می کند. استفاده از تکنیک هایی مانند تطبیق دامنه و یادگیری چند شات می تواند این چالش را به طور موثر برطرف کند.

¹ Quantization

▪ کاوش در سیستم های هوش مصنوعی مشترک

توسعه چارچوب‌هایی که در آن سیستم‌های SOD مبتنی بر DT با سایر ماژول‌های هوش مصنوعی (به عنوان مثال، پردازش زبان طبیعی، کنترل رباتیک، یا تجزیه و تحلیل پیش‌بینی‌کننده) همکاری می‌کنند، می‌تواند راحل‌های یکپارچه را برای کارهای پیچیده و چند وجهی، مانند کاوش مستقل یا یادگیری تطبیقی فعال کند.

▪ بررسی پیامدهای اخلاقی و اجتماعی

با فراگیرتر شدن سیستم‌های SOD، پرداختن به نگرانی‌های اخلاقی بالقوه، از جمله سوگیری در تشخیص برجستگی و پیامدهای تشخیص نادرست در برنامه‌های کاربردی حیاتی مانند مراقبت‌های بهداشتی و امنیت، بسیار مهم است. توسعه سیستم های شفاف و قابل توضیح مبتنی بر DT می‌تواند چنین خطراتی را کاهش دهد.

▪ مقیاس پذیری و یادگیری توزیع شده

پیاده‌سازی چارچوب‌های آموزشی و استنتاج توزیع‌شده برای DT ها می‌تواند مقیاس‌پذیری را افزایش دهد و این مدل را برای مدیریت مجموعه‌های داده در مقیاس بزرگ و برنامه‌های بلادرنگ قابل اجرا کند. این همچنین می‌تواند پذیرش این مدل‌ها را در محیط‌های مبتنی بر ابر یا غیرمتمرکز تسهیل کند.

▪ ترکیب یادگیری تحت نظارت و بدون نظارت

کاوش الگوهای یادگیری ترکیبی که یادگیری نظارت شده را با رویکردهای بدون نظارت یا خود نظارت ترکیب می‌کنند، ممکن است سازگاری مدل را بهبود بخشد و اتکا به داده‌های برچسب‌گذاری شده را کاهش دهد و آن را قادر می‌سازد تا مجموعه داده‌های متنوع را به طور مؤثرتری مدیریت کند.

با پرداختن به این حوزه‌ها، چارچوب پیشنهادی می‌تواند به یک فناوری قوی‌تر، سازگارتر و با کاربرد گسترده‌تر تبدیل شود و تأثیر آن را به طور قابل‌توجهی در صنایع مختلف و حوزه‌های دانشگاهی گسترش دهد.

مراجع

-
- [1] Cheng, M. M., Zhang, G. X., Mitra, N. J., Huang, X., & Hu, S. M. (2011). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 569–582. <https://doi.org/10.1109/TPAMI.2014.2345401>
- [2] Yang, J., Wang, M., Yang, J., & Yuille, A. L. (2013). Correspondence-driven saliency transfer. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1–8. <https://doi.org/10.1109/ICCV.2013.334>
- [3] Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012). Geodesic saliency using background priors. *European Conference on Computer Vision (ECCV)*, 29–42. https://doi.org/10.1007/978-3-642-33712-3_3
- [4] Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., & Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2083–2090. <https://doi.org/10.1109/CVPR.2013.270>
- [5] Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1155–1162. <https://doi.org/10.1109/CVPR.2013.153>
- [6] Zhu, W., Liang, S., Wei, Y., & Sun, J. (2014). Saliency optimization from robust background detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2814–2821. <https://doi.org/10.1109/CVPR.2014.359>
- [7] Hou, Q., Cheng, M. M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4), 815–828. <https://doi.org/10.1109/TPAMI.2018.2844178>
- [8] Luo, Z., Mishra, A. K., Achkar, A., Eichel, J., Li, S., & Jodoin, P. M. (2017). Deep level sets for salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2300–2309. <https://doi.org/10.1109/CVPR.2017.246>
- [9] Wang, L., Wang, W., Lu, H., Zhang, P., & Ruan, X. (2018). Recurrent attentional networks for saliency detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6730–6739. <https://doi.org/10.1109/CVPR.2018.00704>
- [10] Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Deep saliency with encoded low-level distance map and high-level features. *IEEE Transactions on Image Processing*, 26(9), 4206–4217. <https://doi.org/10.1109/TIP.2017.2714793>
- [11] Wang, L., Lu, H., Wang, X., Feng, M., Ding, E., & Ruan, X. (2016). Deep contrast learning for salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 478–487. <https://doi.org/10.1109/CVPR.2016.57>
- [12] Han, J., Shen, X., Sui, X., Liu, D., & Yang, L. (2017). Visual saliency based on multiscale deep features. *IEEE Transactions on Image Processing*, 26(11), 5184–5196. <https://doi.org/10.1109/TIP.2017.2723503>

-
- [13] Liu, N., Han, J., & Yang, M. H. (2016). DHSNet: Deep hierarchical saliency network for salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 678–686. <https://doi.org/10.1109/CVPR.2016.78>
- [14] Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2018). Non-local deep features for salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6609–6617. <https://doi.org/10.1109/CVPR.2018.00692>
- [15] Wang, T., Zeng, Y., Wang, S., & Lu, H. (2017). A stagewise refinement model for detecting salient objects in images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4019–4028. <https://doi.org/10.1109/ICCV.2017.430>
- [16] Li, X., Zhao, L., Wei, L., Wang, M., Wu, F., & Zhuang, Y. (2018). Deep networks for saliency detection via local estimation and global search. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5157–5166. <https://doi.org/10.1109/CVPR.2018.00543>
- [17] Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 202–211. <https://doi.org/10.1109/ICCV.2017.30>
- [18] Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2018). Learning uncertain convolutional features for accurate saliency detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 212–221. <https://doi.org/10.1109/ICCV.2017.33>
- [19] Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1265–1274. <https://doi.org/10.1109/CVPR.2015.7298747>
- [20] Zhang, S., Hu, J., Li, C., Cheng, M. M., & Torr, P. H. (2022). Generative transformer for accurate and reliable salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11238–11247. <https://doi.org/10.1109/CVPR.2022.01101>
- [21] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 15084–15097.
- [22] Jiang, H., Wang, J., Yuan, Z., Zheng, N., Li, S., & Lou, Y. (2013). Salient object detection driven by fixation prediction. *IEEE Transactions on Image Processing*, 22(10), 4318–4331. <https://doi.org/10.1109/TIP.2013.2269905>
- [23] Achanta, R., Hemami, S., Estrada, F., & Süsstrunk, S. (2009). Frequency-tuned salient region detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1597–1604. <https://doi.org/10.1109/CVPR.2009.5206596>
- [24] Borji, A., Cheng, M. M., Jiang, H., & Li, J. (2015). Salient object detection techniques in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4), 697–713. <https://doi.org/10.1109/TPAMI.2014.2359672>

-
- [25] Shi, J., Yan, Q., Xu, L., & Jia, J. (2016). Salient object detection: A discriminative regional feature integration approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2083–2092. <https://doi.org/10.1109/CVPR.2016.229>
- [26] Alexe, B., Deselaers, T., & Ferrari, V. (2010). Learning to detect a salient object. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 150–157. <https://doi.org/10.1109/CVPR.2010.5540226>
- [27] Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., & Wu, F. (2018). Salient object detection: A survey. *Artificial Intelligence Review*, 50(1), 31–66. <https://doi.org/10.1007/s10462-017-9588-5>
- [28] Zhao, R., Ouyang, W., Li, H., Wang, X., & Tang, X. (2019). Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10), 3209–3232. <https://doi.org/10.1109/TNNLS.2019.2929770>
- [29] Yang, W., Zou, C., & Sun, X. (2020). MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 99–110. <https://doi.org/10.1016/j.isprsjprs.2020.06.005>
- [30] Wei, Y., Wen, F., Zhu, W., & Sun, J. (2017). Learning to detect salient objects with image-level supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 136–145. <https://doi.org/10.1109/CVPR.2017.178>
- [31] Zhu, W., Liang, S., Wei, Y., & Sun, J. (2014). Saliency optimization from robust background detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2814–2821. <https://doi.org/10.1109/CVPR.2014.359>
- [32] Liu, Z., Jiang, S., Wei, Y., Zhao, H., Lu, J., & Li, Z. (2021). Visual saliency transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4722–4731. <https://doi.org/10.1109/CVPR.2021.00469>
- [33] Zhang, P., Wang, H., & Lu, H. (2020). Texture-guided saliency distilling for unsupervised salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12234–12243. <https://doi.org/10.1109/CVPR.2020.01256>
- [34] Zhang, T., Li, Y., Xu, X., Yang, Y., & Shen, C. (2022). Recurrent multi-scale transformer for high-resolution salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12035–12044. <https://doi.org/10.1109/CVPR.2022.01267>
- [35] Wang, R., Chen, J., Liu, X., & Yang, X. (2021). Dataset enhancement with instance-level augmentations. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2363–2372. <https://doi.org/10.1109/ICCV.2021.00238>
- [36] Zhao, J., Pang, Y., Zhang, L., Luo, J., Han, J., & Yang, X. (2022). A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection, and video salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12167–12176. <https://doi.org/10.1109/CVPR.2022.01269>

-
- [37] Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., & Jagersand, M. (2019). BASNet: Boundary-aware salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7479–7489. <https://doi.org/10.1109/CVPR.2019.00766>
- [38] Caron, M., Misra, I., Bojanowski, P., Mairal, J., & Joulin, A. (2021). Self-supervised transformers for unsupervised object discovery using normalized cut. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 287–296. <https://doi.org/10.1109/ICCV.2021.00036>
- [39] Wang, Y., Li, Z., Zhang, H., Wu, C., & Zhu, X. (2021). Saliency DETR: Enhancing detection transformer with hierarchical saliency filtering refinement. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4419–4428. <https://doi.org/10.1109/ICCV.2021.00444>
- [40] Zhu, X., Zhang, H., & Yang, Z. (2021). P2T: Pyramid pooling transformer for scene understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8195–8204. <https://doi.org/10.1109/ICCV.2021.00808>
- [41] Li, L., Xu, Y., & Gong, Y. (2021). An energy-based prior for generative saliency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3844–3853. <https://doi.org/10.1109/CVPR46437.2021.00380>
- [42] Li, S., Zhang, Y., & Zhou, X. (2022). Generative transformer for accurate and reliable salient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11321–11330. <https://doi.org/10.1109/CVPR52688.2022.01110>
- [43] Liu, Y., Zhang, J., & Wang, Y. (2020). Advancing saliency ranking with human fixations: Dataset models and benchmarks. *IEEE Transactions on Image Processing*, 29, 1825–1838. <https://doi.org/10.1109/TIP.2020.2973505>
- [44] Zhou, X., Yang, L., & Li, X. (2022). Unifying global-local representations in salient object detection with transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2085–2094. <https://doi.org/10.1109/CVPR52688.2022.00859>
- [45] hang, Z., Li, Y., & Yang, J. (2022). SAP-DETR: Bridging the gap between salient points and queries-based transformer detector for fast model convergence. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2313–2322. <https://doi.org/10.1109/ICCV48922.2021.00232>
- [46] Liu, L., Li, H., & Yang, Z. (2021). Pyramidal attention for saliency detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12021–12030. <https://doi.org/10.1109/CVPR46437.2021.01184>
- [47] Zhang, L., Wang, Z., & Li, J. (2020). Scene context-aware salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6), 1748–1759. <https://doi.org/10.1109/TCSVT.2019.2909271>
- [48] Lajkó, M., & Csuvik, V. (Year). Towards JavaScript program repair with Generative Pre-trained Transformer (GPT-2). *Conference/Journal Name, Volume(Issue), Page Range*. University of Szeged. DOI/Publisher URL.

فهرست واژگان

شماره	واژه ها	معانی
1	Accuracy	دقت
2	Saliency Object Detection	تشخیص قسمت های برجسته ی شی
3	Transformer Decision	مبدل تصمیم
4	Reinforcement Learning	یادگیری تقویتی
5	Sequational Decision	تصمیم گیری متوالی
6	Deep Learning	یادگیری عمیق
7	Recall	دقت
8	Computer vision	بینایی کامپیوتر
9	Convolutional neural networks	شبکه های عصبی کانولوشنی
10	F ₁ Score	F ₁ امتیاز
11	Mean Square Error	میانگین خطای مطلق
12	Intersection on Unity	تقاطع روی واحد
13	Inference Time	زمان استنتاج
14	Segmentation Algorithm	الگوریتم تقسیم بندی
15	Hierarchical strategy	استراتژی سلسله مراتبی
16	Normalization	حاشیه نویسی
17	Dense layers	لایه های متراکم
18	Decoder	رمزگشا
19	Encoder	رمزگذار
20	State	حالت
21	Action	عمل
22	Reward	پاداش
23	Return-To-Go	بازخورد
24	Input Dimension	ابعاد ورودی

25	Self-Attention mechanism	مکانیزم خود توجهی
26	Attention Head	سر توجه
27	Mean Absolute Error	خطای مطلق میانگین
28	Sequence Length	طول دنباله
29	Batch size	اندازه دسته
30	Epoch	دوره
31	Ground Truth	حقیقت پایه
32	Token-to-Token Vision Transformer	مبدل بینایی توکن به توکن
33	Patch-Task-Attention	توجه وصله - وظیفه
34	Binary Cross Entropy	آنترپی متقاطع دودویی
35	Learning rate	نرخ یادگیری
36	Weight decay	کاهش وزن
37	Data Augmentation	افزایش داده
38	Positional Embedding	جاسازی مکانی
39	Mask	ماسک
40	Multi-Head Attention	توجه چند سر
41	Patch	بخش - وصله
42	Bounding Box	جعبه محدود کننده
43	Pre-trained model	مدل از پیش آموزش دیده شده
44	Object Label	برچسب شی
45	Cross Attention	توجه متقاطع
46	Output and Multi-Head Attention	خروجی و توجه چند سری
47	Feed-Forward Layer	لایه پیش خور
48	Add & Norm Layer	لایه جمع و نرمال سازی
49	Object Label	برچسب شی ء
50	Cross-Attention	توجه متقاطع

51	Token	توکن
52	Backbone	ستون فقرات
53	Spatial / Temporal	مکانی / زمانی
54	TimeStep	زمانبندی
55	Structure Measure	معیار ساختاری
56	Harmonic Mean of Precision and Recall	میانگین هارمونیک دقت و بازیابی

Family Name: Rafiei		Name: Ainaz
This Title: Saliency Object detection With Decision Transformer		
Supervisor: Dr.Pedram Salehpoor		
Advisor: Dr.Farshi		
Degree: Master Of Science		Major: ComputerEngineering
Field: Artificial Intelligence and Robotics		
University: University of Tabriz		Faculty: Electrical & Computer Engineering
Graduation Date: 2025.24.01		Page: 89
Key Words: Saliency Object detection, Decision Transformer, Reinforcement Learning		
<p>Abstract</p> <p>Salient object detection in images is a significant topic in computer vision with broad applications such as object recognition, target tracking, and image analysis. This thesis investigates and develops a Transformer-based model called the Decision Transformer to perform salient object detection tasks on the DUTS dataset. The primary goal of this research is to leverage the capabilities of transformers in processing complex data and integrating temporal and visual features to enhance the accuracy and efficiency of salient object detection. The DUTS dataset, one of the most extensive datasets in this field, comprises 10,572 images for training and 5,019 images for testing. In this study, 8,442 images were used for training, 2,111 for validation, and 5,019 for testing. The Decision Transformer, originally designed for reinforcement learning tasks, has been repurposed in this research for salient object detection in images. This model combines features extracted from images using the Vision Transformer (ViT) with temporal information. The architecture of the model includes components such as self-attention layers, feedforward neural networks, and activation functions, enabling it to process multimodal data and learn complex sequences effectively.</p> <p>To evaluate the proposed model, two additional methods were analyzed. The first, Visual Saliency Transformer, is a transformer-based approach that employs multi-head self-attention mechanisms to accurately detect visual saliencies. By focusing on precise representations of both local and global image features, this model significantly improves salient object detection results. The second, Texture-guided Saliency Distilling for Unsupervised Salient Object Detection, utilizes texture-based guidance to distill saliencies in an unsupervised manner. This approach leverages texture information to produce more refined saliency maps. The experimental results demonstrate that the Decision Transformer outperforms the other two models in metrics such as S-measure and F-measure. This research highlights the capabilities of transformer architectures in addressing complex computer vision problems and shows that the Decision Transformer can serve as a powerful tool for salient object detection. Given these findings, this study represents a significant step in expanding the application of transformers to other computer vision tasks.</p>		



University of Tabriz

Faculty of Electrical & Computer Engineering

Department of Computer Engineering

Title

Saliency Object Detection With Decision Transformer

Supervisor

Dr. Pedram Salehpour

Advisor

Dr. Farshi

Researcher

Ainaz Rafiei