ارائه الگوریتمی برای تولید دادگان در ردگیری چند هدفه مبتنی بر یادگیری عمیق

کوروش داداش تبار احمدی [*,۱]

۱- استادیار دانشگاه صنعتی مالک اشتر

چکیده

امروزه با ارائه روش‌های جدید در حوزه هوش مصنوعی نیاز به دادگان آموزشی برای این روش ها در حال افزایش است. بنابراین نیاز جوامع علمی این حوزه به دادگان مرتبط هر لحظه در حال رشد می‌باشد. از این رو محققان همواره درحال کار بر روی روش هایی برای تولید دستی و یا اتوماتیک این دادگان هستند . از طرفی بهبود روش های موجود هر حوزه بعنوان یک بحث همیشگی در جوامع علمی مطرح می باشد. در این مقاله ، ابتدا روشی کاملا خودکار برای افزایش تعداد دادگان مورد استفاده در روش های مبتنی بر ، بخش‌بندی تصویر و ردگیری چند هدفه براساس دادگان موجود ارائه می شود و سپس با ارائه رویکردی نوین با عنوان MOTSNet به بهبود دقت الگوریتم های پیشین می پردازند.

کلمات کلیدی: دادگان آموزشی، ردگیری اشیا، بخش‌بندی تصویر، تصویربرداری نوری

۱. مقدمه

ردگیری اشیا در ویدیو بدلیل عوامل مختلف همانند حرکت دوربین، تغییر شکل شی و تار شدن شی بعلت حرکات دوربین همواره یکی از پرچالش ترین حوزه های بینایی کامپیوتر طی سال های اخیر می باشد . همانطور که در مقاله [2] اشاره شده است از روش های مثل Frag track و Hog-LBP Detector میتوان بعنوان معروف ترین روش قبل از یادگیری عمیق نام برد اما پس ارائه روش های مبتنی بر یادگیری عمیق الگوریتم های دقیق و سریع تر از موارد فوق بوجود آمد منتها این روش ها نیز همواره دارای مشکلاتی در ارائه دقت و سرعت مناسب هستند و در اغلب آن ها با افزایش سرعت، دقت کاهش می یابد و یا بالعکس. بنابراین همواره محقیق در پی راهبردی جهت حل این دسته مشکلات می باشند. حال اگر بخواهیم به جای یک شی چند شی را در تصویر دنبال کنیم مشکلات تشدید می شوند و نیازمند روش های قدرتمندتری برای حل این چنین مسائل خواهیم بود که در این مقاله راهکاری نوین جهت بهبود دقت الگوریتم های ردگیری چند شی در ویدیو بصورت آنلاین مبتنی بر یادگیری عمیق ارائه شده است .

بطورکلی ردگیری چند شی در ویدیو را می توان در دو دسته آنلاین و آفلاین مورد بررسی قرار داد. در چارچوب های آنلاین که روش استفاده شده در مقاله [1] میباشد،عموما از فریم t و t-1 جهت ردگیری اشیای موجود در ویدیو استفاده می کنند. چالش اصلی روش های مبتنی بر چارچوب آنلاین همواره ارائه روش های قدرتمند جهت اتصال حرکت هر شی یافته شده در یک فریم به همان شی در فریم بعد می باشد [3]. از این رو در این مقاله روشی جهت بهبود دقت ردگیری چند شی در

* **Corresponding author:** استادیار مجتمع دانشگاهی برق و کامپیوتر
Email: dadashtabar@mut.ac.ir

ویدیو مبتنی بر DN-MOT یا Tracking-by-Detection ارائه می دهد . همچنین در این مقاله [1] با استفاده از یک شبکه Region Segmentation عملیات بخش بندی اشیا را در ویدیو انجام می دهد که در نتیجه یک روش نوین ردگیری به همراه ارائه ماسک هر شی را تحت عنوان MOTS بوجود می آورد . در این روش با بکارگیری instance segmentation از مشکل همپوشانی bounding box ها جلوگیری میکنند که خود موجود افزایش دقت است در این رویکرد به ساخت representative embedding vector برای هر شی جهت ردگیری می پردازند . ردگیری در این روش بصورت یک تابع نگاشت خطی (LAP) فرموله می شود که نهایتا فرایند ردگیری را دقیق تر و سریع تر می کند.

اولین بار در سال ۲۰۱۹ مقاله [4] ردگیری چند شی در ویدیو را با instance segmentation ترکیب کرد این روش ها که Multi Object Tracking and Segmentation یا به اختصار MOTS نامیده میشوند. بنابر نوین بودن چارچوب به طبع با کمبود دادگان آموزشی روبرو هستند . علیرغم وجود دادگان های همانند DAVISو، YouTube-VIS همواره با کمبود داده برای این روش ها مواجه ایم از طرفی تولید دادگان بصورت دستی و یا نیمه خودکار به شدت زمانبر می باشد لذا باید راهکاری جهت تولید دادگان بصورت خودکار ارائه شود. در این مقاله با ارائه روشی کاملا اتوماتیک که از مفاهیم Optical imaging بهره میگیرد توانستند دیتاست های موجود برای موضوع MOT از جمله KITTI Raw و MOT CHALENGE را به دیتاستی مناسب برای موضوع MOTS تبدیل کنند.

در بخش دوم این مقاله به مرور روش های گذشته خواهیم پرداخت سپس در بخش سوم به شرح سامانه پیشنهادی مقاله می پردازیم و نهایتا در بخش های چهارم و پنجم به ارائه گزارشی از عملکرد مقاله برروی دادگان مختلف و جمع بندی پرداخته می شود در انتها نیز به ایده پیشنهادی جهت بهبود این مقاله بیان می شود.

۲. کارهای مرتبط

باتوجه به اینکه MOTS یکی از حوزه های بسیار جدید در بینایی کامپیوتر میباشد لذا پژوهش های پیشین این حوزه بیشتر مربوط به عناوین MOT و VOS هستند. در این مقاله با توجه به اینکه سه وظیفه را برای روش پیشنهادی خود در نظر گرفته اند بر همین اساس برای هر تسک به طور مجزا پژوهش های پیشین را مورد بررسی قرار داده اند لذا در این مقاله نیز به همین شکل به شرح پیشینه پژوهش پرداخته خواهد شد.

۲-۱-ترکیب حرکت و مفهوم : در مقاله [5] با ترکیب semantic segmentation با شارهای نوری پیش بینی شده به ایجاد یک ground truth برای موضوع آموزش مدل خود پرداخته است و همچنین در مقاله [6] از Mask-RCNN بعنوان یک استخراج کننده ویژگی ، جهت پیش بینی Instance segmentation فریم بعدی استفاده کرده است .

۲-۲- تولید نیمه اتوماتیک دادگان : بسیاری از دادگان های موجود برای چالش های VOS و MOT بصورت دستی حاشیه نویسی شده اند این دیتاست ها هزینه بسیار زیاد و همچنین زمان بالایی را نیاز دارند.در [7] اگرچه عملیات تخصیص ماسک بصورت دستی انجام می شود ولی با بکارگیری همبستگی زمانی بین فریم های پشت سر هم در استراتژی Skip-frame به تولید دادگان می پردازند. همچنین در مقاله [8] با استفاده از آموزش یک شبکه VOS به بخش بندی فریم ها می پردازد اما نهایتا بخش ها بصورتی دستی باید حاشیه نویسی شوند.

۲-۳-روش های مبتنی بر یادگیری عمیق برای MOT و MOTS : برای MOT روش های بسیار گوناگونی ارائه شده است در مقاله [3] این روش ها به سه دسته DN-MOT، DM-MOT وGN-MOT تقسیم می شوند در این مقاله تمرکز بر روش های مبتنی بر DN-MOT و یا Detection-by-tracking می باشد برای MOTS روش های محدودتری وجود

دارد که Mask R-CNN میتوان بعنوان یکی از معروف ترین روش های این حوزه اشاره کرد همچنین در [9] روشی بعنوان CAMOT ارائه شده است. همچنین در [10] روشی مبتنی بر شبکه های SIAMESE ارائه شده اند

۳. طراحی و معماری سامانه پیشنهادی

این قسمت ابتدا به شرح روش تولید دادگان آموزشی در این مقاله خواهیم پرداخت سپس به شرح مدل جدید ارایه شده در [1] می پردازیم.

۳-۱-سامانه تولید دادگان :در این مقاله با استفاده از دادگان KITTI Raw به شرح مدل خود که شامل دو مرحله کلی تولید نمونه های اشیای موجود در هر فریم ویدیو و همچنین در مرحله دوم به تولید نمونه های قابل ردگیری با استفاده از شار نوری می پردازد . قابل ذکر است که دادگان KITTI Raw شامل ۱۴۲ ویدیو (سری تصویر) می باشد.

۳-۱-۱- تولید نمونه های اشیای موجود در هر فریم ویدیو : در این قسمت ابتدا با استفاده از روش Seamless Scene Segmentation که از ResNext101-32*8 بهره می برد و همچنین برروی دادگان Mapillary Vistas آموزش دیده است به استخراج اشیای موجود (instances) در فریم های مختلف می پردازد. یادآور میشوم که دادگان Mapillary Vistas دارای ۳۷ کلاس شی مختلف می باشد.نهایتا در این قسمت ۱.۲۵ سگمنت مختلف از دادگان KITTI Raw استخراج می شود. در ادامه متغیر های تعریف شده برای این قسمت آورده شده است :

y : ۳۷ کلاس اشیای مختلف متعلق به دادگان Mapillary Vistas

S : مجموعه اشیای یافت شده در هر ویدیو موجود در دادگان KITTI Raw

s : هر سگمنت عضو S

t_s : فریمی که سگمنت s از آن استخراج شده است

y_s : کلاسی که سگمنت sبه آن تعلق دارد

$\emptyset_s(i,j): R^2 \to \{0,1\}$: اگر مقدار این مورد یک باشد یعنی پیکسل متعلق به سگمنت s میباشد.

پس تا این قسمت ما ۱.۲۵ میلیون سگمنت مختلف داریم که میباست مشخص کنیم کدامیک قابل ردگیری هستند که در مرحله بعدی به این موضوع خواهیم پرداخت .

۳-۱-۲- تولید اشیای قابل ردگیری با استفاده از شار نوری : در مرحله قبل سگمنت های موجود در دیتاست استخراج شد حال با استفاده از شار نوری به استخراج اشیای قابل ردگیری میپردازیم.

در ابتدا با اعمال مخزن OpenSfM بر روی دیتاست KITTI Raw ساختار سه بعدی برای هر فریم ویدیو تولید می شود و با در نظر گرفتن دو فریم متوالی شار نوری (optical flow) فریم ها بدست می آید سپس با استفاده از مدل بهبود داده شده [12] که برروی دادگان KITTI Raw از نو آموزش داده شده است یک شبکه شار را بوجود می آید و در نهایت با اجرای شبکه ایجاد شده برروی فریم های متوالی ویدیو به دو فریم متوالی را بصورت پیکسل به پیکسل به هم نگاشت می شود که در نهایت یک بردار نگاشت $\overrightarrow{f_t}$ برای هر فریم t بدست می آید.

برای بدست آوردن اشیای قابل ردگیری در فریم t یک گراف $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{T}_t)$ که راس های آن تمام سگمنت های استخراج شده تا بخش t می باشند و همچنین یال های این گراف نشان دهنده سگمنت های مرتبط تا فریم t را نشان می دهد .برای ساخت این گراف مراحل زیر انجام می شود:

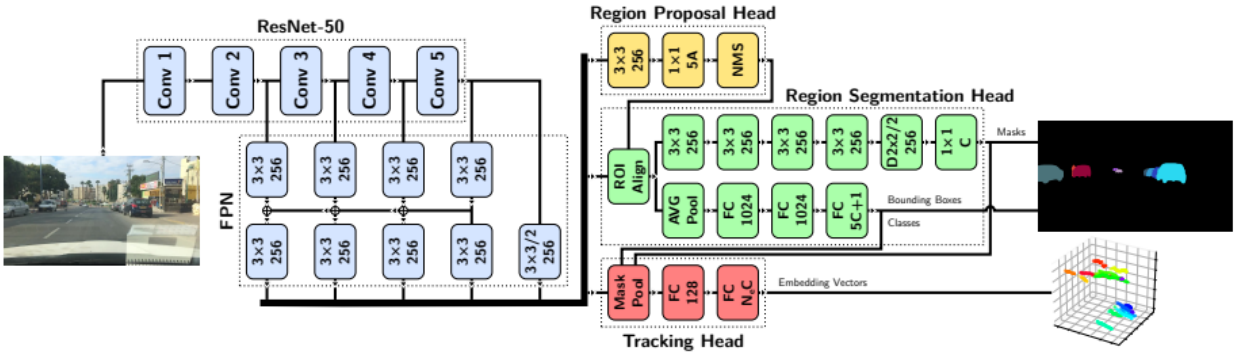
مرحله اول : راس اول شامل تمام سگمنت های استخراج شده از گراف $\mathcal{G}_t$ می باشد. در نتیجه بدست $\mathcal{G}_1 = (S_1, \varphi)$ می آید.

مرحله دوم : هر مرحله t>1 براساس مرحله قبلی و با استفاده از سگمت های استخراج شده در مرحله t و بردار نگاشت $\overrightarrow{f_t}$ بدست می آید.

در توالی ساخت گراف، رئوس جدید که همان اشیای جدید هستند که توالی فریم ها پدیدار میشوند بوجود می آیند این فرایند با اجتماع رئوس گراف در مرحله t-1 و سگمنت های بدست آمده در مرحله S_t به گراف اصلی اضافه می شوند.

برای بدست آوردن مجموعه یال های این گراف یک مسئله نگاشت خطی بین S_t و S_{t-1} با استفاده از الگوریتم مجارستانی حل می شود در آن هدف بیشینه کردن تابع payoff زیر است که نهایتا رئوس دارای بیشترین تشابه به هم متصل می شوند در این تابع payoff با استفاده از IoU بین سگمنت های فریم t-1 و t و همچنین تابع η که نشان دهنده تعلق s_t و s_{t-1} به یک کلاس می باشد ایجاد می شود.مقادیر تابع η بین صفر و بینهایت منفی متغیر است و در صورتی که صفر شود دو سگمنت به یک کلاس تعلق دارند و هر چه این تابع کوچک تر باشد میزان شباهت دو سگمنت کمتر می باشد.

$$\pi(s_{t-1}, s_t) = IoU\left(\phi_s, \phi_{t-1}{}^{\circ}\overrightarrow{f_t}\right) + \eta(s_{t-1}, s_t) \qquad (1)$$



شکل ۱- در معماری سامانه MOTSNet ، شبکه هرم ویژگی و ResNet-50 به رنگ آبی ، RPH به رنگ زرد ، RSH به رنگ سبز و بخش ردگیری به رنگ قرمز آورده شده است.

۲-۳- رویکرد MOTSNet : رویکرد MOTSNET مبتنی بر رویکرد Mask R-CNN ارایه شده است که در ادامه به شرح این رویکرد خواهیم پرداخت.

رویکرد MOTSNet مبتنی بر رویکرد Mask R-CNN طراحی شده است نویسندگان مقاله [1] با اضافه کردن یک ردگیر اشیا (Tracking Head) به Mask R-CNN رویکرد جدید خود را ارائه داده اند در معماری این رویکرد ابتدا یک شبکه هرم ویژگی (FPN) روی شبکه ResNet-50 قرار گرفته است تا برداری های ویژگی را برای ۵ اندازه مختلف از ResNet-50 استخراج کند سپس برداری های بوجود آمده جهت استخراج قاب های اشیا (Bounding Box) به RPN یا Region Proposal Headداده می شوند پس از استخراج، این قاب ها به همراه بردارهای خروجی از FPN ، جهت استخراج ماسک های هر شی برای Region Segmentation Head ارسال خواهند شد و در نهایت پس استخراج ماسک ها ، اطلاعات FPN و RSH در اختیار قسمت ردگیری قرار می گیرند در این قسمت ابتدا توسط لایه Mask Pooling پسزمینه اشیا حذف می شود که در نهایت دو لایه تمام متصل موجود در قسمت ردگیری بتواند فرآیند ردگیری اشیا طی فریم های مختلف را انجام دهند. این فرآیند به این صورت می پذیرد که این دو لایه تمام متصل فضایی را بوجود می آورند که طی آن بردارهای منتصب به یک شی واحد در این فضا طی فریم های متوالی در کنار هم قرار می گیرند. این فرآیند نیز با کمینه سازی تابع هزینه بیان شده در تساوی شماره ۲ صورت می پذیرد.

$$L = L_{TH} + \lambda(L_{RPH} + L_{RSH}) \qquad (2)$$

توابع هزینه L_{RSH} و L_{RPH}در مقاله [11] بطور کامل شرح داده شده اند و تابع هزینه L_{TH} در تساوی ۳ داده شده است.

$$L_{TH} = -\frac{1}{|S_B|}\sum_{s\epsilon S_B}\max\left(\max_{s\epsilon\mu_B(s)}\left\|a_s^{y_s} - a_{\hat{s}}^{y_{\hat{s}}}\right\| - \min_{\hat{s}\epsilon\mu_B(s)}\left\|a_s^{y_s} - a_{\hat{s}}^{y_{\hat{s}}}\right\|\right) - (\beta, 0) \quad (3)$$

$$\hat{s} = s_{t-1}$$
$$s = s_t$$

۴. نتایج و مباحث

در این بخش ابتدا به بررسی عملکرد سامانه ساخت دیتاست جدید خواهیم پرداخت سپس نتایج حاصل از MOTSNet را روی دادگان KITTI MOTS مورد بررسی قرار میدهیم. در این مقالهاز شاخص های ردگیری اشیا MOTS ، MOTSP و sMOTSA استفاده شده است و همچنین برای بررسی دقت کلی شناسایی ماسک ها و قاب های اشیا موجود در فریم های ویدیو شاخص mAP را انتخاب کرده اند.

۴-۱- بررسی عملکرد سامانه ساخت دادگان جدید :برای بررسی عملکرد سامانه ساخت دادگان ارائه شده توسط این مقاله ، ابتدا مدل CAMOT روی دادگان تولید شده توسط سامانه آموزش داده می شود سپس با استفاده از دادگان ارزیابی KITTI MOTS عملکرد سامانه بررسی می شود. دو ردیف ابتدایی جدول شماره ۱ نشان دهنده عملکرد دو مدل متفاوت سامانه پیشنهادی می باشد در مدل اول از شبکه شار نوری HD3 Model zoo استفاده شده است و همچنین از HD3 Model KITTI Sfm برای ساخت شبکه شار نوری مدل دوم استفاده شده است. با مقایسه نتایج موجود در جدول شماره ۱ با مقاله [9] که در این مقاله مدل CAMOT برای آموزش و ارزیابی از KITTI MOTS استفاده کرده است ، می توان به قابل قبول بودن عملکرد سامانه پیشنهادی پی برد.در دو قسمت بعدی جدول، دادگان تولید شده با مدل پیشنهادی MOTSNet مورد ارزیابی قرار گرفته اند در قسمت میانی عملکرد دادگان با بخش ردگیری متفاوت قابل مشاهده است و همچنین در قسمت آخر نیز میتوان عملکرد این سامانه یک بار بصورت پیش آموزش روی ImageNet و یک بار بدون پیش آموزش دید. در همه موارد داده شده، این دیتاست دارای عملکرد بسیار مناسبی می باشد.

۴-۲- بررسی عملکرد MOTSNet : نویسندگان این مقاله برای بررسی عملکرد مدل *MOTSNet* ، ابتدای مدل های پیش آموزش دیده روی دادگان های *Imagenet* ، *Mapillary Vistas* و دادگان تولید شده با سامانه پیشنهادی همین مقاله ، با استفاده از دادگان KITTI MOTS تنظیم میکنند (fine-tune) . سپس MOTSNet را نیز هر بار با یکی از دادگان های معرفی شده آموزش می دهند و سپس با KITTI MOTS تنظیم نهایی می کنند و نتایج حاصل را روی دادگان ارزیابی KITTI MOTS مورد بررسی قرار داده می شود. این نتایج در جدول شماره ۲ آورده شده است.
رویکرد MOTSNet که با استفاده از دادگان های معرفی شده در همین قسمت مورد پیش آموزش قرار گرفته است روی دادگان KITTI MOTS دقت را برای عابرین پیاده حدود ۷.۵ درصد و برای خودروها حدود ۱.۹ درصد نسبت به بهترین روش قبلی مقاله [4] افزایش می دهد.

Method	Pre-training	sMOTSA Car	sMOTSA Ped	MOTSA Car	MOTSA Ped	MOTSP Car	MOTSP Ped	mAP Box	mAP Mask
KITTI Synth (val) + HD^3 [49] model zoo	inference only	65.4	45.7	77.3	66.3	**87.6**	76.6	–	–
KITTI Synth (val) + HD^3, KITTI-SfM	inference only	65.5	45.4	77.4	66.0	**87.6**	76.6	–	–
MOTSNet with:									
AVEBOX+TH	I	73.7	46.4	85.8	62.8	86.7	76.7	57.4	50.9
AVEMSK-TH	I	76.4	44.0	88.5	60.3	86.8	76.6	57.8	51.3
AVEBOX-TH	I	75.4	44.5	87.3	60.8	86.9	76.7	57.5	51.0
KITTI MOTS train sequences only	I	72.6	45.1	84.9	62.9	86.1	75.6	52.5	47.6
MOTSNet	I	77.6	49.1	89.4	65.6	87.1	76.4	58.1	51.8
MOTSNet	I, M	**77.8**	**54.5**	**89.7**	**70.9**	87.1	**78.2**	**60.8**	**54.1**

جدول شماره ۱ – نتایج عملکرد سامانه پیشنهادی ساخت دادگان

Method	Pre-training	sMOTSA Car	sMOTSA Ped	MOTSA Car	MOTSA Ped	MOTSP Car	MOTSP Ped	mAP Box	mAP Mask
TrackR-CNN [43]	I, C, M	76.2	47.1	87.8	65.5	87.2	75.7	–	–
CAMOT [31]	I, C, M	67.4	39.5	78.6	57.6	86.5	73.1	–	–
CIWT [30]	I, C, M	68.1	42.9	79.4	61.0	86.7	75.7	–	–
BeyondPixels [40]	I, C, M	76.9	–	**89.7**	–	86.5	–	–	–
MOTSNet	I	69.0	45.4	78.7	61.8	88.0	76.5	55.2	49.3
	I, M	74.9	53.1	83.9	67.8	89.4	79.4	60.8	54.9
	I, KS	76.4	48.1	86.2	64.3	88.7	77.2	59.7	53.3
	I, M, KS	**78.1**	**54.6**	87.2	**69.3**	**89.6**	**79.7**	**62.4**	**55.7**

جدول شماره ۲- نتایج عملکرد رویکرد MOTSNet

۵. نتیجه گیری و پیشنهاد

در مقاله ارائه شده از مقاله [1] ابتدا سامانه تولید دادگان جدید برای چارچوب MOTS مورد بحث واقع شد. این سامانه ابتکاری برای دادگان فضای خیابان براساس دادگان KITTI Raw عملکرد بسیار مناسبی را ارائه می دهد. سپس روش ابتکاری MOTSNet مورد بررسی قرار گرفت این روش نسبت به بسیاری از روش های مطرح پیشین براساس دادگان KITTI Raw عملکرد مناسب تری می باشد. و همچنین ارائه یک لایه Mask pooling جدید در رویکرد MOTSNetو استفاده ابتکاری از شبکه HD3 Flow Network در سامانه تولید دادگان جدید را میتوان به عنوان ایده های جذاب مقاله [1] در نظر گرفت.

$$\frac{\partial \vec{r}}{\partial \psi_w}+\frac{\partial \vec{r}}{\partial \psi_b}=\frac{1}{\Omega}\vec{V}(\vec{r},t) \quad (1)$$

که در آن ψ_w و ψ_bمتغیرهای ... برای ذره‌ای با سرعت $\vec{V}$در موقعیت $\vec{r}$و زمان t هستند.

روش های ردگیری اشیا مبتنی بر شناسایی شی در هر فریم یا *tracking-by-detection* همواره با مشکل سرعت پایین مواجه هستند در مقاله *[1]*نیز مبتنی بر چارچوب ذکر شده ارائه شده است به احتمال بسیار زیاد دارای سرعت پایینی در ردگیری یک شی می باشد (در نظر بگیرد که در مقاله هیچ صحبتی از موضوع سرعت نشده است و امکان تست و پیاده سازی این شبکه با لپ تاپ شخصی تقریبا غیرممکن است لذا بیان موضوع سرعت پایین تنها براساس رویکردهای مشابه رویکرد مقاله *[1]* بیان شده است) لذا از دید بنده میتوان با اضافه کردن قسمت شناسایی شی (قسمت آبی رنگ تصویر شماره ۱) با مدل ارائه شده در مقاله *[12]* که از قسمت شناسایی شی جدید در تمامی فریم ها عاجز است به رویکردی با سرعت بالاتر و دقت مناسب برسیم.

۱۲. مراجع

[1] L.Porzi, M.Hofinger, I.Ruiz, J.Serrat, S.RotaBulò, P.Kontschieder, Learning Multi-Object Tracking and Segmentation from Automatic Annotations, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[2] K. Rasool Reddy, K. Hari Priya, N. Neelima, Object Detection and Tracking – A Survey, act 2015 International Conference on Computational Intelligence and Communication Networks.

[3] S.Cheng, Y.Xu, X.Zhou, Deep Learning for Multiple Object Tracking: A Survey, Jan 2019 IET Computer Vision

[4] P.Voigtlaender, M. Krause, A. Osep, J.Luiten, B.Balachandar, G.Sekar, A.Geiger,B.Leibe. Mots: Multi-object tracking and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[5] X.Jin, H.Xiao, X.Shen, J.Yang, Z.Lin, Y.Chen, Z.Jie, J.Feng,S.Yan. Predicting scene parsing and motion dynamics in the future. In Neural Information Processing Systems, 2017.

[6] P.Luc, C.Couprie, Y.LeCun, J.Verbeek. Predicting future instance segmentations by forecasting convolutional features. In Proceedings of the European Conference on Computer Vision, 2018.

[7] N.Xu, L.Yang, Y.Fan, D.Yue, Y.Liang, J.Yang, T.Huang. Youtube-vos: A large-scale video object segmentation benchmark. arXiv:1809.03327, 2018.

[8] A.Berg, J.Johnander, F.Durand, J.Ahlberg, M.Felberg. Semi-automatic annotation of objects in visual-thermal video. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[9] A.Osep, W.Mehner, P.Voigtlaender, B.tian L.Track, then decide: Category-agnostic vision based multi-object tracking. In Proceedings of the IEEE International Conference on Robotics and Automation,2018.

[10] B.Li, J.Yan, W.Wu, Z.Zhu, X.Hu. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[11] L.Porzi, S.Rota Bulo, A.Colovic, P.Kontschieder. Seamless scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[12] Q. Wang, L. Zhang, L. Bertinetto, W. Hu and P. H. S. Torr, "Fast Online Object Tracking and Segmentation: A Unifying Approach," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1328-1338.