Tabriz Journal of Electrical Engineering (TJEE), vol. 54, no. 4, December 2024 Serial no. 110
DOI: 10.22034/tjee.2024.59362.4766

Long-term Visual Object Tracking of Arbitrary Objects Based on Switching Between Traditional Method and Deep Learning Technique

Mohammad Amin Bagherzadeh, Hadi Seyedarabi*, Seyed Naser Razavi

Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran
E-mails: amin.bagherzadeh@tabrizu.ac.ir; seyedarabi@tabrizu.ac.ir; n.razavi@tabrizu.ac.ir

Short Abstract
Visual tracking of the arbitrary object is a fundamental and challenging topic in the field of machine vision, which has traditionally been done by considering a model for the target and using the training data of the same video. Most trackers can hardly top the results of the most popular methods when considering real-time and online performance. In this article, a tracker framework based on the Siamese network is presented, which is an online tracker learning and a real-time tracking process, and its name is STD-Siam. Since the Siamese network has limited online training and cannot handle the challenges of tracking for the long term, STD-Siam aims to switch between traditional tracking and deep learning, training both trackers to eliminate the ambiguity between the target and the background in each scenario. First, the training data is generated through the traditional tracker, then these data are expanded with the augmentation technique so that the deep network can be trained well. This method can be executed at a speed of 66 FPS, and compared to the current similar algorithms, despite its simplicity, it can achieve good results and track the target for the long term. This tracking speed is beyond real-time due to the spike detector in the frequency domain, which accurately calculates the selected target candidates and avoids blindly scanning the entire image to reduce the computational burden.

Keywords
Long-term visual tracking, Siamese neural network, deep learning.

1- Short Introduction

Online, real time, and long-term visual object tracking is a fundamental research problem in computer vision, artificial intelligence, and video analytics. Object tracking aims to find the arbitrary object position in the subsequent video frames. Without prior knowledge of the target category, the tracker is set to the first frame with an initial bounding box. The mission of a tracker system is to estimate the target's position in the following video frames, in which often the target is non-rigid. In addition, the bounding box around the target is shown in each frame. Therefore, the visual tracking problem can be considered a combination of classification, estimation, and detection tasks.

2- Proposed Work and Methodology

This paper introduces an end-to-end trainable STD-Siam method for long-term and real-time visual object tracking. The tracker performs the tracking by switching between the two modes of deep learning and the traditional method. At the beginning of the tracking process, the desired target is tracked using a spatio-temporal context, and the labeled data results are sent to the augmentation unit. The defect of low data is solved by using augmentation, and by cut-out the images, it can significantly overcome the challenge of relative occlusion of the target in tracking. After the completion of learning, the deep learning tracker continues the process of tracking the target. In addition, scanning the whole image with the saliency detector uses all the information to detect the target in the image quickly. We further demonstrate the usefulness of our tracker for long-term tracking despite the challenges. STD-Siam also achieved state-of-the-art visual object trackers on large datasets like VOT2019/2020, VOT2018, VOT2016, OTB100, LaSOT, and GOT-10K benchmarks while operating at 66FPS

3- Conclusion

STD-Siam tracker can constantly capture the right arbitrary target. In general, our tracker performs better with a high number of frames because the traditional tracker must be able to prepare several frames in which the object is located and send it to the data augmentation unit. Then, by augmenting the data, the training unit starts working. If the number of frames is too low, the deep learning tracker will not be activated. The results prove that our STD-Siam tracker can handle many complex tracking environments with non-rigid targets. Compared to state-of-the-art trackers, it shows competitive performance and long-term visual tracking of the target. In the future, we intend to increase the accuracy and robustness of tracking by using the integration of radar sensor and camera data to improve the tracker's performance in adverse weather conditions.

4- References

[1] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., and Torr, P.H.: Fully-convolutional Siamese networks for object tracking. in *European Conference on Computer Vision (ECCV Workshops)* (2016).

[2] Henriques, J.F., Caseiro, R., Martins, P., and Batista, J.: High-speed tracking with kernelized correlation filters. in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(3), pp. 583–596 (2015).

رهگیری بصری طولانی مدت اهداف دلخواه بر اساس راه‌گزین بین دو روش رهگیری سنتی و فن یادگیری ژرف

محمدامین باقرزاده
دانشجوی دکتری، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

میرهادی سیدعربی
استاد، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

سید ناصر رضوی
استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

چکیده
رهگیری بصری شی دلخواه یک موضوع اساسی و چالش برانگیز در حوزه بینایی ماشین است که به‌طور سنتی توسط در نظر گرفتن یک مدل برای هدف و با استفاده از داده‌های آموزشی همان ویدیو انجام شده است. اکثر رهگیرها به‌سختی می‌توانند با در نظر گرفتن ویژگی‌های برخط و بی‌درنگ در صدر مقایسه نتایج مشهورترین روش‌ها قرار گیرند. در این مقاله یک چارچوب رهگیری بر اساس شبکه سیامی ارائه شده که یادگیری رهگیر به‌صورت برخط و فرآیند رهگیری بی‌درنگ بوده و نام آن STD-Siam است. از آن‌جا که شبکه سیامی دارای محدودیت آموزش برخط است و مدت طولانی نمی‌تواند چالش‌های موجود در رهگیری را مدیریت کند، هدف STD-Siam از راه‌گزین[1] بین رهگیر سنتی و رهگیر بر مبنای یادگیری ژرف، تعلیم هر دو رهگیر با هدف رفع ابهام بین هدف و پس‌زمینه در هر فرنامه دلخواه است. ابتدا از طریق رهگیر سنتی داده‌های آموزشی تولید شده، سپس این داده‌ها با فن برافزایی گسترش داده می‌شوند تا شبکه ژرف به خوبی آموزش بیند. این روش می‌تواند با سرعت ۶۶ فریم‌درثانیه اجرا شود و نسبت به الگوریتم‌های مشابه فعلی با وجود سادگی آن نتایج خوبی را به‌دست آورد و به‌صورت طولانی مدت هدف را رهگیری کند. این سرعت رهگیری فراتر از بی‌درنگ (بیش از ۳۰ فریم در ثانیه) به‌واسطه آشکارساز برجستگی در حوزه فرکانس است که نامزدهای انتخابی هدف به‌طور دقیق محاسبه شده و از روبش کل تصویر به‌صورت کورکورانه جلوگیری می‌شود تا بار محاسباتی کاهش یابد.

کلمات کلیدی
رهگیری بصری طولانی مدت، شبکه پیچشی سیامی، یادگیری ژرف.

نام نویسنده مسئول: دکتر میرهادی سیدعربی
ایمیل نویسنده مسئول: seyedarabi @tabrizu.ac.ir

تاریخ ارسال مقاله: ۱۴۰۲/۰۹/۱۵
تاریخ(های) اصلاح مقاله: ۱۴۰۲/۱۱/۲۴
تاریخ پذیرش مقاله: ۱۴۰۲/۱۲/۲۰

۱- مقدمه

هدف از رهگیری بصری پیدا کردن موقعیت شی دلخواه در توالی فریم‌های ویدیو است. رهگیر بدون داشتن اطلاعات از دسته‌بندی هدف، در اولین فریم با یک جعبه‌محاطی اولیه مقداردهی می‌شود. وظیفه رهگیر تخمین مکان و ناحیه هدف در فریم‌های پیش‌رو است و در هر فریم جعبه‌محاطی دور هدف نشان داده می‌شود. بنابراین، مسئله رهگیری را می‌توان ترکیبی از دو وظیفه دسته‌بندی و تخمین تلقی کرد. دسته‌بندی تمایل دارد هدف انتخاب شده در اولین فریم را با سایر فریم‌ها مقایسه کرده و مکان هدف را تعیین نماید. وظیفه دوم مربوط به برآورد وضعیت هدف با یک جعبه‌محاطی است که می تواند از لحاظ اندازه ثابت یا متغیر باشد. از مهم‌ترین کاربردهای رهگیری بصری می‌توان کنترل ترافیک [۱]، تعامل انسان و کامپیوتر [۲]، نظارت بصری [۳]، واقعیت افزوده [۴]، وسایل نقیله خودران و سامانه‌های رهگیری هوایی [۵، ۶] را نام برد. به نظر می‌رسد کلید رهگیری بصری با توانایی تعقیب هدف به‌صورت طولانی مدت، در پیدا کردن ویژگی‌های هدف توسط دسته‌بندی کننده‌های عمومی و تبعیض‌پذیر است. توانایی تبعیض‌پذیری این امکان را فراهم می‌کند که رهگیر بین هدف و پس‌زمینه تبعیض قائل شود. زیرا ممکن است اطراف هدف، اشیا مشابه وجود داشته باشد و رهگیر دچار شکست شود. با توجه به تغییرات ظاهری هدف و تداخلات پس‌زمینه، داشتن این دو خصوصیت نیاز به یادگیری برخط در حین رهگیری دارد زیرا محیط رهگیری همیشه پس‌زمینه ساده‌ای ندارد. دو استراتژی برای رهگیری وجود دارد: حالت اول، استفاده از دسته‌بندی کننده کلاسیک با به‌روزرسانی مدل ظاهری هدف [۷] است تا به‌صورت برخط [۸، ۹] محتمل‌ترین نامزد نمونه را به‌عنوان هدف تحت رهگیری در فریم بعدی انتخاب نماید. روش‌های رهگیری هدف با چنین دسته‌بندی کننده‌هایی توانسته نتایج خوبی را با مدل‌های از قبل آموزش دیده شده بدست آورد [۱۰-۱۳]، که امروزه این کار توسط تنظیم دقیق و برخط شبکه‌های ژرف توسعه یافته است. شبکه‌های‌عصبی پیچشی[2] (CNNs) توانسته‌اند در بسیاری از

[1] Switch
[1] Convolutional Neural Neworks

حوزه‌های بینایی ماشین به‌خصوص رهگیری هدف پیشرفت چشمگیری داشته باشند [۱۴-۱۷]. ، برخی از روش‌های رهگیری با شبکه‌های ژرف خاص مثل MDNet[3] [۱۸] وجود دارند که به‌دلیل پیچیدگی شبکه بسیار کند هستند ولی از دقت بالایی برخوردارند. البته این بدان معنا نیست که پیچیده کردن شبکه و افزایش لایه‌ها و ژرفای آن منجر به افزایش دقت می‌شود. مشابه به MDNet، روش‌های رهگیری STCT [۱۹] و DeepSRDCF [۲۰] با سرعت ۲ فریم در ثانیه اجرا می‌شوند ولی دارای دقت بالایی هستند.

استراتژی دوم، رهگیری بر اساس تطبیق الگوی هدف با وصله‌های نامزد در فریم‌های آتی است که این روش‌ها معمولا از به‌روزرسانی برخط جهت آموزش شبکه استفاده نمی‌کنند. مزیت قابل توجه این روش‌های بر پایه CNN سرعت بی‌درنگ آن‌هاست که برخی جهت افزایش قدرت تعمیم‌پذیری از شبکه‌های ژرف استفاده می‌کنند [۲۱، ۲۲]. یک الگوریتم رهگیری موفق که بستر بسیاری از روش‌ها و تحقیقات نیز بوده، SiamFC [۱۴] است که از یک تابع تطبیق کلی برای تغییرات آنلاین هدف استفاده می‌کند. مهم‌ترین دلیل ناتوانی در رقابت با رهگیرهایی مثل MDNet این است که تغییرات ظاهری اهداف و کلاترها را نمی‌تواند به‌صورت برخط ثبت کند و از یک مدل از قبل آموزش دیده، برای رهگیری استفاده می‌کند. این باعث می‌شود که از لحاظ دقت رهگیری نسبت به سایر روش‌ها [۲۳] شاهد یک شکاف باشیم که در پایگاه داده‌هایی نظیر OTB2015 [۲۴] قابل مشاهده است.

عملکرد مناسب شبکه سیامی باعث شده تا رهگیرهای مختلفی بر مبنای این شبکه توسط محققان توسعه داده شود [۱۴، ۲۵-۲۷]. این روش‌های رهگیری بصری با محاسبه همبستگی متقابل بین نواحی کاندیدای هدف و ویژگی‌های معرف الگوی هدف، مکان هدف را در فریم‌های آتی پیدا می‌کنند. نکته قابل توجه این است که همه این رهگیرها معماری مشابه به شبکه AlexNet [۲۸] را ایجاد کرده‌اند. حتی تلاشهایی برای ایجاد یک معماری پیچیده‌تر مانند ResNet [۲۹] وجود داشته که کارایی قابل توجهی نداشته‌اند. اگرچه رهگیرهای سیامی عملکرد فوق‌العاده‌ای به ویژه در سرعت داشته‌اند، ولی عدم توانایی آن‌ها در آموزش شبکه به‌صورت برخط باعث می‌شود که وجود اشیا مشابه در همسایگی هدف یا تغییرات ظاهری زیاد، رهگیری را دچار شکست کند که در مورد SiamFC نیز این‌چنین است. راه‌حل ساده برای این مشکل جایگزینی قالب ظاهری هدف با نتایج رهگیری به‌دست آمده از فریم‌های قبل است [۲۰] که باعث می‌شود تغییرات ظاهری هدف در طول زمان در نظر گرفته شود. اما می‌دانیم که همیشه نتایج رهگیری درست نیست. اگر در فریم قبل، مکان هدف به اشتباه تشخیص داده شود، در فریم بعدی این مکان جهت تطبیق استفاده می‌شود و رهگیر را دچار شکست خواهد کرد. رهگیر STMTrack [۳۰] یکی دیگر از روش‌های مبتنی بر چارچوب سیامی است که از یک شبکه حافظه فضا – زمانی استفاده کرده تا بتواند اطلاعات هدف را در طول رهگیری ذخیره و استفاده نماید. رهگیر LightTrack [۳۱] با هدف اجرای الگوریتم بر بستر پردازنده‌های تلفن همراه یک شبکه با تعداد پارامتر کم جهت کاهش بار محاسباتی را ارائه می‌دهد.

در این مقاله، رویکرد اصلی استفاده از شبکه‌ای است که در حین رهگیری با داده‌های جدید آموزش ببیند. از آنجا که ممکن است هدف در هر موقعیتی با تغییرات ظاهری شدید، انسداد نسبی و کامل و تغییرات نور ظاهر شود، در کنار داده‌های جدید ارسالی به واحد آموزش، داده‌های برافزوده[4] نیز تولید می‌شود تا به کمک آن‌ها ضعف کمبود داده در آموزش برخط شبکه برطرف شده و فرآیند رهگیری هدف به‌خوبی انجام شود. رهگیر پیشنهادی می‌تواند به‌صورت برخط آموزش ببیند. این امر با آموزش برون‌خط شبکه با داده‌های آموزشی و سپس به‌روزرسانی آن با داده‌های ورودی و تصاویر جدید امکان‌پذیر است. مهم‌ترین ویژگی‌های روش پیشنهادی را می‌توان به‌صورت زیر خلاصه کرد.

۱- یک رهگیر سنتی قابل اطمینان جهت آغاز فرآیند رهگیری در اولین فریم پیشنهاد شده تا با انتخاب کاربر، جعبه‌محاطی اطراف هدف مورد نظر ترسیم شود. وظیفه این رهگیر تعقیب هدف و همزمان با آن، تولید داده‌های آموزشی برای رهگیر مبتنی بر یادگیری ژرف است.

۲-در حین رهگیری برخط، واحد تولید تصاویر آموزشی و برافزوده، داده‌های جدیدی از روی تصاویر رهگیر سنتی تولید می‌کند که هدف این بخش مقابله با چالش‌های پیش‌روی رهگیری است و به عمومی‌سازی آن کمک می‌کند. با افزایش تعداد داده‌ها یادگیری شبکه ژرف با حداقل تعداد داده‌ها امکان‌پذیر می‌شود.

۳-آموزش شبکه سیامی و راه‌گزین کردن از حالت سنتی به یادگیری ژرف جهت ادامه فرآیند رهگیری از طریق شبکه‌ای که با داده‌های جدید آموزش دیده، انجام می شود.

۴-روش ارائه شده توانسته در مقایسه با جدیدترین روش‌های مشابه، نتایج خوبی را در پایگاه‌داده‌های مطرح رهگیری [۳۲-۳۴] به‌دست آورد. با مقدار دهی جعبه‌محاطی در اولین فریم و عدم تغییر آن، رهگیر توانسته با سرعت بی‌درنگ (بیش از ۳۰ فریم در ثانیه) در توالی‌های ویدیویی طولانی مدت هدف را رهگیری کند.

۲- پیشینه تحقیق

رهگیرهای مبتنی بر مدل هدف معمولا برای رهگیری یک کلاس خاص از قبل آموزش دیده شده، طراحی شده‌اند [۳۵-۳۷]. به‌عنوان مثال، اگر در یک سیستم نیاز به رهگیری یکی از کلاس‌های خودرو، عابرپیاده، و یا توپ فوتبال باشد، می‌توان یکی از این مدل‌ها را به آشکارساز آموزش داد. محدودیت این نوع رهگیرها در آموزش برخط، رهگیری فقط یک کلاس است که نمی‌توان برای اشیای جدید بدون آموزش از آن استفاده کرد. مکان‌یابی هدف تحت رهگیری به‌طور قابل توجهی بستگی به توانایی تبعیض‌پذیری بین هدف و پس‌زمینه دارد. این ویژگی در رهگیرهای تبعیض‌پذیر بر مبنای تطبیق الگو، به‌عنوان فیلتر همبستگی افتراقی (DCF)[5] مورد مطالعه قرار گرفته است که یکی از روش‌های محبوب برای رهگیری هدف است. در مقاله [۳۸]، MOSSE[6] به‌عنوان رهگیر بصری برمبنای DCF با سرعت ۶۶۹ فریم درثانیه معرفی شد که با استفاده از ویژگی‌های تک کاناله فضای رنگی خاکستری اجرا می‌شود. رهگیرهای KCF [۳۹] و RLSC [۴۰] از دیگر روش‌های مبتنی بر DCF هستند که از ماتریس گردشی برای رهگیری هدف استفاده کرده و عملیات پردازشی را در فضای فرکانسی با تبدیل فوریه انجام می‌دهند. همچنین از آزمون چند مقیاسی برای تخمین مقیاس هدف استفاده کرده‌اند که این کار با جمع‌آوری یک دسته کوچک از تصاویر مقیاس شده به روش روبش نواحی مختلف تصویر و استفاده از دسته‌بندی کننده جهت پیدا کردن بالاترین امتیاز انجام می شود. Hu و همکارانش [۴۱] روشی با عنوان DCFNet ارائه داده‌اند که بر مبنای شبکه سیامی است و با استفاده از یک معماری سبک وزن جهت یادگیری ویژگی‌های پیچشی و استفاده از یک لایه فیلتر همبستگی اضافه شده در شبکه سیامی، رهگیری هدف را با سرعت ۶۰ فریم درثانیه انجام می‌دهد. رهگیر ATOM [۴۲] از دو وظیفه دسته‌بندی هدف و تخمین هدف استفاده می‌کند. دسته‌بندی سعی می‌کند هدف و پس‌زمینه را از همدیگر متمایز کند و تخمین هدف با رویکرد

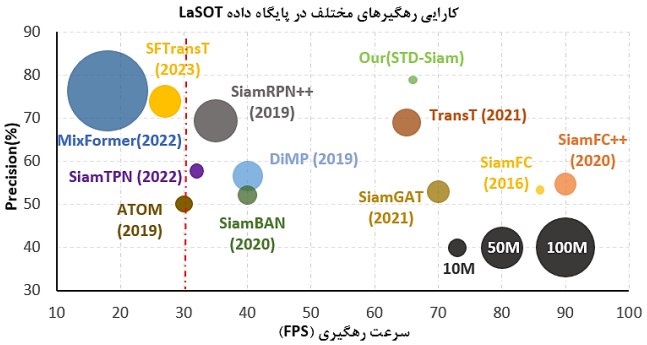
[3] Multi-Domain Convolutional Neural Networks
[2] Augment
[5] Discriminative Correlation Filter
[6] Minimum Output Sum of Squared Error

حداکثرسازی همپوشانی انجام می‌شود که در آن یک ماژول شبکه مسئول افزایش همپوشانی بین هدف و جعبه‌محاطی، با آموزش آنلاین شبکه است. این روش دارای پارامترهای زیادی است که بار محاسباتی را سنگین می‌کند.

از آنجایی که سهم اصلی این مقاله رهگیری بر اساس شبکه سیامی است، در ادامه به‌طور خلاصه به معرفی رهگیرها و پیشینه اخیر در روش رهگیری مبتنی بر سیامی پرداخته می‌شود. رهگیری بصری یک مسئله بینایی ماشینی حیاتی است که به‌عنوان تطبیق شباهت بین یک مدل خاص و نامزدها در یک فریم ویدیو مدل‌سازی می‌شود. اخیراً، با استفاده گسترده از شبکه‌های عصبی پیچشی ژرف، گروهی از رهگیرهای بصری مبتنی بر شبکه سیامی پدید آمده‌اند که تطبیق شباهت را بر اساس ویژگی‌های استخراج‌شده DCNN انجام می‌دهند. پیشگام این نوع رهگیر، روش کاملاً کانولوشنی سیامی (SiamFC) است. SiamFC [۱۴] با استفاده از معماری AlexNet [۲۸] ویژگی‌های شبکه عصبی پیچشی ژرف را از ناحیه هدف و منطقه جستجو استخراج می‌کند. سپس با استفاده از همبستگی دو نقشه ویژگی، یک نقشه پاسخ را تشکیل می‌دهد. بالاترین مقدار نقشه پاسخ، مکان هدف درتصویر است. مزیت این روش عدم نیاز به آموزش برخط است. البته این تا زمانی است که چالش‌های اساسی مانند تغییرات ظاهری هدف، تغییر نور، حرکت‌سریع، درهم‌تنیدگی پس‌زمینه وجود نداشته باشند. بنابراین، برای کاربردهای نسبتا ساده مفید است. به‌منظور بهبود عملکرد شبکه سیامی، روش SINT[7] [۴۳] از شبکه سیامی برای رهگیری بصری استفاده می‌کند. این استراتژی به دلیل استفاده از شبکه ژرف برای استخراج ویژگی و جریان نوری برای نمونه برداری بسیار دقیق‌تر است اما بسیار کندتر از SiamFC است. رهگیر GOTURN [۴۴] از یک پنجره‌کشویی با نمونه‌برداری تصادفی برای تولید تصاویر نامزد استفاده می‌کند. برخلاف سایر مدل‌های تطبیق الگو این روش با مقایسه مختصات جعبه‌محاطی مرزی هدف در فریم فعلی و قبلی، مختصات هدف را محاسبه می‌کند.

بر اساس معیارهای ارزیابی OTB2015 [۲۴]، روش‌های مبتنی بر سیامی بدون شک در سرعت و دقت رهگیری عملکرد خوبی دارند. حتی بهترین روش‌ها مانند SiamRPN [۴۵] هنوز با پیشرفت‌های روز فاصله زیادی دارند. در SiamRPN، یک شبکه پیشنهاد منطقه، با دو بخش دسته‌بندی و رگرسیون پس از شبکه سیامی برای استخراج ویژگی پیشنهاد شده است که به صورت برون‌خط آموزش داده می‌شود و رهگیری شی با سرعت ۱۶۰ فریم در ثانیه انجام می‌شود. این ساختار بیشتر برای تشخیص شی و تخمین مقیاس هدف مناسب است اما در برابر تغییرات ظاهری هدف مقاوم نیست. رهگیر D3S [۴۶] سعی کرده این نقص را با راه‌اندازی دو مدل تطبیقی و متمایز و تغییر ناپذیر نسبت به تغییرات گسترده برای هدف برطرف کند که با استفاده از قطعه‌بندی هدف انجام می‌شود. این رهگیر با سرعت کمتری نسبت به بی‌درنگ کار می‌کند. رهگیر SiamRCNN [۴۷] بر اساس معماری شبکه سیامی، رهگیری را با تشخیص مجدد برای مقایسه مناطق پیشنهادی با الگوی هدف فراهم می‌کند. در پژوهشی دیگر، رهگیر SiamRPN++ [۴۸] از یک شبکه پیشنهادی ژرف برای رهگیری هدف جهت تطبیق بیشتر با اندازه جسم و شناسایی آن استفاده می‌کند. روش رهگیری SiamMask [۴۹] از یک افت قطعه‌بندی دودویی و یک محدوده قطعه‌بندی برای تعیین جعبه‌محاطی مرزی هدف استفاده می‌کند. در روش رهگیری DiMP [۵۰] جهت رفع محدودیت‌های شبکه سیامی که نمی‌تواند به‌صورت برخط آموزش بیند، این روش یک راه‌حل ذخیره تغییرات ظاهری هدف و پس‌زمینه را ارائه می‌دهد. رهگیر SiamRTU [۵۱] نیز سعی کرده تا با استفاده از ذخیره‌سازی تغییرات و ویژگی‌های ظاهری هدف به‌صورت برخط بخش یادگیری شبکه را تقویت کند که باز نتوانسته موفقیت چندانی کسب کند. در پژوهشی دیگر SiamTC [۵۲]، یک شبکه سیامی با قابلیت شناخت هدف ارائه شده که با به‌کارگیری یک بلوک توجه شناختی هدف بین شاخه‌های الگو و جستجو، سعی در افزایش قدرت رهگیری داشته‌اند. رهگیر SFTransT [۵۳] با الهام از شبکه سیامی و استفاده از اطلاعات فضایی و فرکانسی هدف و همچنین یادگیری ویژگی‌های تعاملی بین شاخه‌های جستجو و قالب توانسته تا حدودی بهبود در نتایج رهگیری داشته باشد که در یکی از پایگاه‌داده‌های مورد استفاده در این پژوهش نتایج برتر را به‌دست آورده است.

از آنجایی‌که قصد مقایسه با سایر رهگیرها را داریم، در اینجا به اختصار به چند ردیاب پیشرفته اشاره می‌شود. روش TrDiMP [۵۴] از یک شبکه ترانسفورمر برای بهبود ویژگی‌های جستجو و الگوی استخراج شده هدف استفاده می‌کند تا شناسایی شی با یک دسته‌بندی متمایز کننده به‌کار گرفته شود. در رهگیر MixFormer [۵۵] یک چارچوب ردیابی فشرده بر اساس شبکه ترنسفورمر است که از ماژول توجه ترکیبی جهت ادغام داده‌های هدف و استخراج ویژگی‌ها در حین تعقیب هدف استفاده می‌کند. رهگیر TransT [۵۶] مستقیماً ویژگی‌های متمایز کننده را با استفاده از ماژول ادغام‌ساز ویژگی استخراج می‌کند که از توجه به خود و توجه متقاطع استفاده می‌کند. الگوریتم KeepTrack [۵۷] رهگیری هدف را با استفاده از آموزش شبکه عصبی گراف انجام می‌دهد و از مناطق مجاور هدف برای انتخاب نامزدهای هدف استفاده می‌کند. روش NeighborTrack [۵۸] رهگیری را با فرمول‌بندی مسئله تطبیق بین نامزد هدف و همسایگی آن انجام می‌دهد. اطلاعات همسایگی نامزد هدف برای تصحیح خطای رهگیری در حین انسداد استفاده می‌شود. خلاصه‌ای از نتایج ارزیابی رهگیرهای مشهور بر اساس معیارهای مختلف در شکل ۱ قابل مشاهده است.



شکل ۱: مقایسه رهگیرهای مختلف بر اساس دقت، سرعت رهگیری و تعداد پارامترهای شبکه که قطر هر دایره است. خط عمودی نشانگر شرایط مرزی سرعت بی‌درنگ است که رهگیرهای سمت راست خط، فراتر از بی‌درنگ اجرا می‌شوند. بهترین نتایج در بخش سمت راست و بالای تصویر قرار دارند.

رهگیر پیشنهادی در این مقاله با استراتژی‌های فوق در موارد زیر متفاوت است. اولاً، راه‌گزینی بین دو رهگیر سنتی و یادگیری ژرف، این دو را قادر می‌سازد تا به طور موثر با هم کار کنند و اطمینان حاصل شود که رهگیرها به طور مستقل نتایج ردیابی را ارائه می‌دهند. رهگیر مبتنی بر روش سیامی قبلاً در روش پیشنهادی آموزش دیده است. اطلاعات مکان هدف محاسبه شده توسط رهگیر سنتی برای به‌روز رسانی و آموزش شبکه CNN به‌صورت برخط استفاده می‌شود تا توانایی رهگیر برای رسیدن به اهداف با تغییرات بصری قابل توجه، بهبود یابد. ثانیاً، مکان هدف در فریم‌های آتی ناشناخته است. اگر لازم

[7] Siamese INstance search Tracker

باشد کل تصویر روبش شود، زمان همپوشانی افزایش می‌یابد و الگوریتم به کندی اجرا می شود. در این تحقیق، یک مکانیزم تشخیص هدف ارائه شده که با محدود کردن مناطق جستجو از کل تصویر به مکان خاصی از تصویر، زمان روبش را سریع‌تر می‌کند.

۳-روش پیشنهادی

در روش پیشنهادی این مقاله، شروع تعقیب هدف به‌عهده یک رهگیر سنتی است تا فرآیند به‌صورت هم‌زمان و برخط انجام شود و مکان هدف در فریم‌های متوالی مشخص شود. رهگیر مبتنی بر یادگیری ژرف به‌صورت برون‌خط آموزش داده شده و برای افزایش استحکام ردیابی طولانی مدت، با داده‌های جدید رهگیر سنتی نیز دوباره آموزش داده می‌شود. رویکرد پیشنهادی در بخش ۳-۳ شرح داده خواهد شد.

۳-۱-رهگیر یادگیری ژرف

وظیفه رهگیری مبتنی بر شبکه سیامی را می‌توان به عنوان یک مسئله یادگیری همبستگی و شباهت مطرح کرد. شبکه سیامی به‌صورت برون‌خط توسط توالی‌های ویدیویی آموزش داده می‌شود و به صورت برخط ارزیابی می‌شود. معماری شبکه SiamFC از دو شاخه CNN تشکیل شده است. یکی شاخه الگو نامیده می‌شود و دیگری شاخه تشخیص یا شاخه جستجو، که پارامترهای مشابهی دارند. ناحیه هدف یا وصله الگو که با z مشخص می شود از فریم اولیه برداشته می‌شود. فریم فعلی به‌عنوان یک تصویر جستجو در نظر گرفته شده که با نماد x نشان داده می‌شود و رهگیر باید شبیه‌ترین وصله یا پچ را پیدا کند. هدف از این کار، محاسبه همبستگی متقابل بین وصله هدف و وصله جستجو از فریم x و همچنین آموزش شبکه برای یادگیری یک تابع تطبیق تشابه کلی است. مشابه به روش SiamFC [۱۴]، امتیاز شباهت با رابطه (۱) به‌دست می آید:

$$SimScore\big(z,x(i)\big)=\sum_{(m,n)\in S}\varphi(z)_{(m,n)}\times\varphi(x(i))_{(m,n)}, \qquad (۱)$$

که در آن $x(i)$ i-امین وصله یا نمونه x با اندازه مشابه z است. نقشه‌های ویژگی $\varphi(x)$ و $\varphi(z)$ وصله‌های ورودی به شبکه AlexNet هستند و φ فرآیند استخراج ویژگی را نشان می‌دهد. $\varphi(z)_{(m,n)}$ بردار مقادیر پیکسلی است که در مختصات $(m.n)$ روی نقشه ویژگی z قرار دارد. S کل مساحت هر نقشه ویژگی را نشان می‌دهد. بنابراین، روبش $\varphi(x)$ با $\varphi(z)$ یک نقشه امتیازی ایجاد می‌کند که هر شباهت بین منطقه x و الگوی z را نشان می‌دهد و حداکثر مقدار در هر ناحیه نشان‌دهنده مکان هدف است. با توجه به آموزش CNN به‌صورت تکثیربه‌عقب[8]، SiamFC با یک روش پیش‌آموزشی برون‌خط با تابع تلفات لجستیک آموزش می‌بینند:

$$l(g,r)=\log(1+\exp(-gr)), \qquad (۲)$$

که در آن g معرف مرجع درستی برچسب داده‌ها است و $g\in\{\pm1\}$ ، r نمایانگر نقشه امتیاز برای جفت داده‌های الگو و نامزد است که در لایه آخر شبکه بدست می‌آید. بنابراین، میانگین تلفات نهایی نقشه امتیاز به شرح زیر است:

$$L(g,r)=\frac{1}{|D|}\sum_{u\in D}l(g[u],r[u]). \qquad (۳)$$

$L(g,r)$ میانگین تلفات نقشه امتیاز است، که در آن D کل مساحت نقشه امتیاز بوده و برچسب‌های درستی $g[u]\in\{\pm1\}$ برای هر مکان $u\in D$ در نقشه امتیاز را نشان می‌دهد. پارامترهای θ که مربوط به شبکه ژرف می‌باشد با استفاده از روش آماری گرادیان‌نزولی (SGD) برای مسئله $\arg\min_{\theta}L(g,r(z,x;\theta))$ به‌دست آمده است.

۳-۲-رهگیر سنتی مبتنی بر دسته‌بندی حداقل مربعات منظم شده

در این قسمت به بررسی دسته‌بندی کننده بر اساس حداقل مربعات منظم شده[9] (RLSC) می‌پردازیم. بردار داده‌های آموزشی را می‌توان به‌صورت $x_i\in\mathbb{R}^n, i=1,\dots,\ell$ نوشت که در دو کلاس y_i معرفی و به صورت زوج $\{(x_i,y_i)\}_{i=1}^{\ell}$ نمایش داده می‌شوند. ℓ تعداد داده‌ها و y_iها برچسب کلاس داده‌ها هستند. هدف، آموزش تابع تصمیم‌گیری f است که باید داده‌های دو کلاس را متمایز کند تا این دو به‌خوبی از هم جدا شوند. تابع f باید به‌صورت عمومی باشد تا از بیش‌برازش جلوگیری کند. یک طبقه بندی کننده‌ی خطی بر اساس تابع تصمیم‌گیری خطی به صورت $f(x)=w^Tx+b$ نمایش داده می‌شود که w بردار وزن‌ها و $w^Tx=\sum_i w_ix_i$ است. b نیز قسمت بایاس بوده و یک عدد ثابتی است که در اینجا صفر در نظر گرفته شده است.

مهم‌ترین بخش در تکنیک‌های یادگیری ماشین، مسئله‌ی کمینه سازی تابع خطرپذیری به صورت عمومی است. با توجه به[۳۹]، مسئله‌ی حداقل مربعات منظم شده را می‌توان به‌صورت رابطه (۴) نوشت:

$$minimize\ H(w):=R_{emp}+\lambda\,\psi(w), \qquad (۴)$$

که $R_{emp}(w)=\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,f(x_i))$ می‌باشد و همان خطرپذیری تجربی است. $\psi(w)$ که در ادامه تعریف می‌شود بخش افزایش حاشیه تصمیم‌گیری را به‌عهده دارد با پارامتر کنترل بیش برازش λ که مصالحه بین بخش کاهش خطرپذیری و افزایش تعمیم‌پذیری است. برای دستیابی به یک تقریب خوب جهت پاسخ دسته‌بندی کننده، تابع زیان به‌صورت مربعی $V(y_i,f(x_i))=(y_i-<w,x_i>)^2$ در نظر گرفته شده که $<.,.>$ ضرب نقطه‌ای است و $<w,x_i>$ همان تابع تصمیم‌گیری خطی است . با توجه به خواص قدرتمند تابع هسته که برای جامعه‌ی یادگیری ماشین شناخته شده است، داده‌های غیرخطی ورودی می‌توانند در فضایی با ابعاد بالاتر توسط ابر صفحه‌ی جدا کننده بهینه از هم جدا شده و دسته‌بندی شوند. تابع هسته بر اساس یک نگاشت است که به‌صورت $\Phi: x\in\mathcal{X}\rightarrow\mathcal{H}: \Phi(x)=k(x,.)$ تعریف می‌شود. این نگاشت برای داده‌های فضای $\mathcal{X}$ است به فضای $\mathcal{H}$، که فضای $\mathcal{H}$ فضای ویژگی نامیده می‌شود و $k(x,.)$ معرف تابع‌هسته است. به علاوه، یکی از شیوه‌های نمایش تابع هسته $k(x,x')$ به‌صورت $k(x,x')=<\Phi(x),\Phi(x')>=\Phi^T(x)\,\Phi(x)$ می‌باشد. بردار وزن‌ها نیز به‌صورت $w=\sum_{i=1}^{\ell}c_i\,\Phi(x_i)$ نوشته می‌شود که c_i ضرایب این بردار است. با در نظر گرفتن $\psi(w)=\lambda\|w\|^2$ [۳۹] و با توجه به رابطه (۴)، این مسئله RLSC را می‌توان به‌صورت رابطه (۵) نوشت:

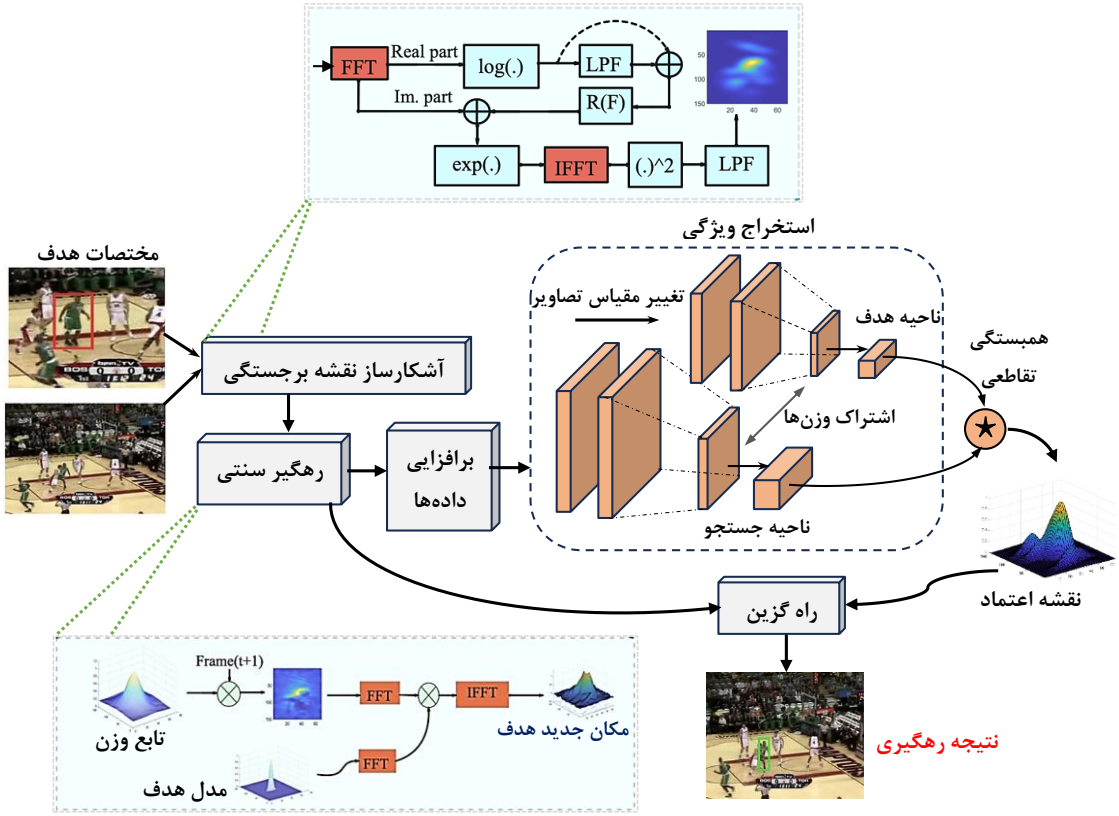
$$\min_{w}\ \frac{1}{\ell}\sum_{i=1}^{\ell}(f(x_i)-y_i)^2+\lambda\|w\|^2, \qquad (۵)$$

وقتی‌که $\|w\|^2=c^TKc$ و K یک ماتریس $\ell\times\ell$ است که (i,j) امین درایه آن برابر است با $k(x_i,x_j)$ و $w=\sum_{i=1}^{\ell}c_i\,k(x,x_i)$ می‌باشد. مسئله آموزش و آشکارسازی در فرآیند رهگیری معادل است با کمینه سازی $H(w)$ و پیدا کردن c_i ها، که با در نظر گرفتن تابع زیان به‌صورت مربعی و مشتق نسبت به c ، پس از اعمال روابط خواهیم داشت:

$$c=\frac{y}{(K+\lambda I)}, \qquad (۶)$$

با توجه به اینکه ماتریس هسته از نوع ماتریس‌گردشی[10] است، امکان نمایش c در فضای فوریه [۳۹] فراهم می‌شود و به‌صورت رابطه (۷) خواهد بود که در آن از عملیات محاسباتی عنصر-گرا استفاده شده تا پیچیدگی محاسباتی کاهش یابد و نتایج به‌صورت یکپارچه و بدون حلقه‌های تکرار شونده به دست آیند.

[8] Backbone
[9] Regularized Least-Square Classification
[10] Circulant Matrix



شکل ۲: معماری رهگیر STD-Siam

جهت محاسبه بردار وزن‌ها که به‌طور ضمنی توسط بردار c نشان داده می‌شود و عناصر آن شامل همه ضرایب c_i است داریم:

$$c = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(y)}{\mathcal{F}(K)+\lambda}\right), \quad (۷)$$

که در این جا $\mathcal{F}$ نماد تبدیل فوریه و $\mathcal{F}^{-1}$ نماد عکس تبدیل فوریه است. مدل گرافیکی هدف شامل رابطه فضایی بین جسم و بافت محلی آن است. به حداکثر رساندن پاسخ دسته‌بندی کننده باعث می‌شود که مرکز هدف در فریم فعلی مشخص گردد. رابطه کلی برای محاسبه پاسخ دسته‌بندی کننده به ازای تک ورودی z به‌صورت $y' = \sum_i c_i\ k(x_i, z)$ است که پیاده‌سازی این رابطه بر اساس محاسبات FFT است و بر پایه‌ی محاسبات عنصر-گرا که با نماد $\odot$ مشخص شده، تمامی پاسخ‌ها در کل تصویر، در یک عملیات کلی به دست می‌آید و داریم:

$$\hat{y} = \mathcal{F}^{-1}(\mathcal{F}(k)\odot\mathcal{F}(c)) . \quad (۸)$$

۳-۳-معماری روش STD-Siam

در این تحقیق یک معماری آموزش‌پذیر برای حل مشکل رهگیری شی به‌صورت بی‌درنگ در یک ویدیو پیشنهاد شده است. در فریم اول ویدیو، محل شی مورد نظر مشخص می‌شود و سپس رهگیر سنتی در تمامی فریم‌های آتی، مکان هدف را تعیین می‌کند تا مجموعه فریم‌های برچسب‌دار آموزشی جهت یادگیری شبکه تامین شود. با جمع‌آوری این داده‌ها، واحد برافزایی از لحاظ کمی تعداد داده‌ها را افزایش می‌دهد تا در یادگیری شبکه ژرف، آموزش به‌خوبی انجام پذیرد. از طرفی شبکه سیامی جهت تشکیل نقشه امتیاز نیاز به روبش کل تصویر دارد که با آشکارساز برجستگی محدوده روبش را معطوف به مکان‌هایی کرده‌ایم که امکان وجود هدف در آن نقاط زیادتر از سایر نواحی تصویر است. معماری کامل الگوریتم رهگیر در شکل ۲ نشان داده شده که با استفاده از رهگیر سنتی در کنار رهگیر مبتنی بر یادگیری ژرف، بهره‌برداری کامل از بعد زمانی و پس‌زمینه هدف انجام شده است. آموزش برخط باعث شده تا تمایز بین هدف و پس‌زمینه افزایش یابد و حتی در انسداد نسبی و کامل، رهگیر بتواند هدف را شناسایی کند.

برای تعیین مکان هدف، رهگیر باید مختصات هدف در فریم قبلی را بداند. از آنجا که اشیا تمایل به حرکت در فضا دارند، یک حدس خوب این است که از ناحیه‌ای در اطراف آخرین مکان هدف استفاده شود. اگر این منطقه جستجو خیلی بزرگ در نظر گرفته شود، در صورت وجود هدف، احتمال یافتن آن بیشتر خواهد بود. با این حال نکته مهم این است که هر چه منطقه روبش بزرگتر باشد، فرآیند جستجو پیچیده‌تر و طولانی‌تر می‌شود. در تحقیق حاضر، از آشکارساز نقشه‌برجستگی برای یافتن تمام مکان‌های نامزد هدف استفاده شده است. علاوه بر گستردگی ناحیه جستجو، سرعت رهگیری کاهش نیافته زیرا این کار از تحلیل فوریه برای تشخیص سریع نقشه‌برجستگی استفاده می‌کند. مزیت گسترش ناحیه روبش، رهگیری اهداف با سرعت بالا و غلبه بر چالش انسداد نسبی یا طولانی مدت است. برای شناسایی اهداف در تصویر، ترجیح این است که آشکارساز عمومی باشد و به ویژگی خاصی مربوط نباشد. مانند پژوهش [۵۹]، بدون داشتن دانش آماری قبلی از صحنه، افزونگی در تصاویر حذف می‌شود تا اشیا یا نامزد مورد نظر نمایان شوند. نقشه برجستگی تصویر ورودی را می‌توان به‌صورت زیر تعریف کرد:

$$S(t) = g(t) * \mathcal{F}^{-1}[\exp(\Re(t) + \mathcal{P}(t))]^2, \quad (۹)$$

که $g(t)$ معرف فیلترگاوسی، $\mathcal{P}(t)$ طیف فاز تصویر ورودی، $*$ نماد عملیات پیچشی و باقیمانده طیفی $\Re(t)$ که معادل است با اختلاف لگاریتم طیفی سیگنال از میانگین طیفی، به‌صورت رابطه (۱۰) محاسبه می‌شود:

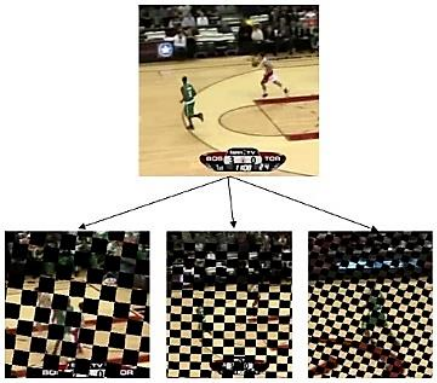
$$\Re(t) = log(Real(\mathcal{F}[z(t)])) - h_n(t) * log(Real(\mathcal{F}[z(t)])), \quad (۱۰)$$

که در آن $h_n(t)$ یک فیلتر میانگین‌گیری با ماتریس $n \times n$ است، $z(t)$ تصویر

ورودی و $Real(\mathcal{F}[z(t)]$ بخش حقیقی تبدیل‌فوریه است.

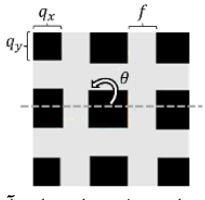
۳-۴-واحد تولید تصاویر آموزشی برافزوده

با توجه به این‌که یکی از مهم‌ترین اهداف این تحقیق، آموزش برخط شبکه و استفاده از رهگیر به‌صورت بی‌درنگ است، از ۲۰۰ فریم رهگیر سنتی برای آموزش شبکه استفاده می‌شود که با استفاده از فن برافزایی داده‌ها، تعداد این تصاویر به ۱۰۰۰ فریم افزایش می‌یابد. استفاده از این تعداد داده آموزشی به طور موثر قابلیت تعمیم‌پذیری شبکه را برای داده‌های جدید افزایش می‌دهد و یادگیری شبکه بدون ایجاد وقفه در عملکرد رهگیر انجام می‌شود. به‌منظور کمک به حل چالش‌های انسداد نسبی و طولانی‌مدت، تغییر روشنایی، تغییرات ظاهری و مقیاس هدف، تصاویر تولید در واحد برافزایی داده‌ها، به سایر تصاویر با برچسب‌های آموزشی که رهگیر سنتی تولید می‌کند اضافه می‌شوند. تولید این داده‌ها با افزودن نویز به تصاویر، چرخش، تغییر مقیاس هدف و حذف نواحی مختلف از آن انجام می‌شود. تمام داده‌های برافزوده به‌صورت تصادفی تولید می‌شوند. یکی از فرآیندهای ضروری برای تولید داده‌های برافزوده، روش حذف نواحی مختلف از تصویر است که باعث می‌شود رهگیر در مواجهه با انسداد هدف قوی‌تر شود. برش یک فن ساده افزایش داده است که برخی از پیکسل‌های مناطقی از تصویر را به‌صورت تصادفی حذف می‌کند و مقدار آن را صفر قرار می‌دهد. با الهام از روش [۶۰] ، پس از بهنجارسازی تصویر، تابع برش و حذف نواحی مختلف تصویر را می‌توان به‌صورت $I_{tr}^{'} = I_{tr} \times BM$ نوشت. وقتی که $I_{tr} \in R^{H\times W\times C}$ نمایانگر تصویر ورودی است، $BM \in \{0,1\}^{H\times W}$ نقاب دودویی است که مکان پیکسل‌هایی که قرار است صفر شوند در آن ذخیره می‌شود و $I_{tr}^{'} \in R^{H\times W\times C}$ تصویر نهایی پس از اعمال نقاب است. اگر $BM_{i,j} = 0$ باشد مقدار پیکسل (i,j) در تصویر ورودی صفر خواهد شد و در غیر این‌صورت آن مقدار بدون تغییر خواهد ماند. با توجه به آزمایش‌ها، اندازه نواحی برش داده شده به‌طور قابل‌توجهی بر عملکرد این روش تأثیر می‌گذارد و شکل نواحی تأثیر کمتری دارد .بنابراین، شکل BM به‌عنوان یک شبکه در نظر گرفته شد، همان‌طور که در شکل ۳ نشان داده شده است.



شکل ۳: اعمال نقاب تولید داده‌های آموزشی روی تصاویر

چهار پارامتر (θ, f, q_x, q_y) برای تشکیل ساختار نقاب استفاده شده که بخشی از هر نقاب در شکل ۴ نشان داده شده است.



شکل ۴: نقاب تولید داده‌های آموزشی

نرخ چرخش ماسک BM برابر است با $\theta = rand(0, 30^{\circ})$. فاصله بین دو ناحیه صفر متوالی $f = rand(f_{min}, f_{max})$ است. f_{min} و f_{max} بر اساس جعبه‌محاطی هدف تعیین می‌شوند تا در صورت بزرگ بودن ناحیه صفر، هدف مسدود نشود. q_x و q_y طول و عرض هر ناحیه صفر از نقاب است. $q_x,\ q_y = rand(0, sum(BB)/2)$ و $sum(BB)$ تعداد کل پیکسل‌ها در جعبه‌محاطی هدف مورد رهگیری است. K تعداد واحدهای با مساحت صفر است که بر اساس اندازه تصویر و مقادیر تصادفی به‌دست آمده برای f ، q_x و q_y محاسبه می‌شود. با توجه به آزمایش‌های انجام شده، انتخاب K باید به‌گونه‌ای باشد که تعداد پیکسل‌های حذف شده از تصویر کمتر از ۴۰ درصد باشد. در غیر این‌صورت، احتمال از دست دادن اطلاعات هدف افزایش می‌یابد.

۳-۵-جزئیات پیاده‌سازی

همان‌طور که گفته شد، شبکه با استفاده از تصاویر خروجی از رهگیر سنتی آموزش داده می‌شود. علاوه بر این، شبکه با استفاده از داده‌های آموزشی از قبل آموزش داده شده است. بنابراین، پارامترهای شبکه با داده‌های آموزشی در هر فرآیند رهگیری به‌روز رسانی می شود تا قدرت رهگیر در مواجهه با سناریوهای جدید افزایش یابد. در پیاده‌سازی روش پیشنهادی، مشابه به SiamFC، جهت آموزش و ارزیابی شبکه وصله هدف z دارای ابعاد ۳×۱۲۷×۱۲۷ و منطقه جستجوی x دارای اندازه ۳×۲۵۵×۲۵۵ است. از گرادیان‌نزولی‌تصادفی (SGD) برای آموزش شبکه از ابتدا استفاده شده که ضریب فروپاشی وزن روی ۰/۰۰۰۵ و تکانه ۰/۹ قرار داده شده است. نرخ یادگیری در فضای لگاریتمی در هر دوره از 10^{-2} به 10^{-5} کاهش می‌یابد و ۳۰ دوره با اندازه بسته ۸ اجرا شده است. اگر اندازه جعبه‌محاطی هدف به‌صورت (w,h) نشان داده شود، وصله الگو با مرکزیت قاب قبلی با اندازه $A \times A$ برش داده شده و به‌صورت $(w+p)\times(h+p) = A^2$ است که در آن حاشیه محتوا $p = \frac{w+h}{2}$ است ودر مراحل بعد به ۱۲۷×۱۲۷ تغییر اندازه داده می‌شود. پارامتر کنترل بیش برازش λ برابر مقدار 10^{-4} در نظر گرفته شده و مقیاس هدف با مقیاس جدید و هموارسازی نمایی $s_t = (1-\vartheta)\times s_{t-1} + \vartheta s_{new}$ به‌روز رسانی می‌شود که در این رابطه s مقیاس هدف و $\vartheta = 0.4$ ضریب به‌روز رسانی است.

رهگیر STD-Siam پس از دریافت داده‌های آموزشی در حین ردیابی شی، شبکه ژرف سیامی را آموزش می‌دهد. کنترل عملیات رهگیری (تغییر از سنتی به ردیابی مبتنی بر یادگیری ژرف) با تکمیل جمع آوری داده‌های آموزشی شروع می‌شود. به‌جای روبش کورکورانه کل تصویر جهت مشخص نمودن نامزدهای مستعد هدف، آشکارساز نقشه‌برجستگی به‌سرعت آن را استخراج می‌کند. با انجام این کار، تمام نقاط نامزد هدف ارزیابی می‌شوند. هر چه افزونگی تصویر بیشتر باشد، نقاط نامزد هدف کمتر و عملیات رهگیری سریع‌تر است. روش پیشنهادی با استفاده از PyTorch در حالی که با سرعت بی‌درنگ اجرا می‌شود، پیاده‌سازی شده است. همه آزمایش‌ها روی رایانه با پردازنده Intel Corei7با فرکانس ۲/۶۰ گیگاهرتز، ۳۲ گیگابایت رم و یک پردازنده گرافیکی NVIDIA GeForce GTX 1080 انجام شده است. در مقابل، ۱/۶ثانیه برای بارگذاری شبکه روی GPU و مقداردهی اولیه مورد نیاز است.

برای بررسی کارایی روش پیشنهادی، مطالعه فرسایشی روی چند موضوع مختلف از جمله کاهش تعداد لایه‌های شبکه، برافزایی و تقویت داده‌ها و تعداد دوره‌های آموزشی صورت گرفته است. از آنجا که هدف ما در روش پیشنهادی، رهگیری با قید فراتر از سرعت بی‌درنگ بوده، افزایش لایه‌های شبکه موجب اندکی افزایش دقت می‌شود ولی سرعت رهگیری به شدت کاهش می‌یابد. نکته‌ای که در رهگیر مبتنی بر شبکه سیامی موجب شکست آن می‌شود، عدم توانایی واحد یادگیری به‌صورت برخط است. یکی از موثرترین کارهایی که موجب افزایش دقت رهگیری در این پژوهش شده، برافزایی به روش برش است. در این روش علاوه بر برش، که نباید بیش از ۴۰ درصد تصویر از دست رود، چرخش نقاب نیز انجام می‌شود که در صورت بیشتر شدن میزان چرخش از ۳۰

درجه، شکست رهگیر در اغلب موارد حتمی و عملکرد آن کوتاه مدت می‌شود. برافزایی داده به روش برش نقش انسداد نسبی را دارد و چرخش تا حدودی به توانایی رهگیر در تعقیب اهداف غیر صلب می‌افزاید. تعداد دوره‌های آموزشی نهایتا ۳۰ دوره می‌باشد که بیشتر از آن نهایتا بهبود ۱ تا ۲ درصد در پی دارد که ترجیح بر آن بوده تا سرعت رهگیر بالاتر بماند. زیرا رهگیر در حین عملکرد در حال سنتی، با تکمیل داده‌ها، فرآیند آموزش شبکه ژرف را شروع می‌شود.

۴-نتایج ارزیابی

در این بخش، رهگیر پیشنهادی در ردیابی استاندارد اشیا برای معیارهای رهگیری کوتاه مدت و بلند مدت با چندین رهگیر مبتنی بر CNN مقایسه و ارزیابی می‌شود. پایگاه داده رهگیری بصری (VOT2019/2020) [۶۱، ۶۲] یکی از مهم‌ترین رویدادهای سالانه در زمینه رهگیری شی است که یک پایگاه داده بلند مدت با ۵۰ دنباله مختلف و بیش از ۲۱۵ هزار فریم است. رهگیر باید بتواند تعداد فریم‌های باقیمانده و مختصات مکانی جعبه محاطی هدف را اعلام کند. عملکرد رهگیر با Rob.[11] (کسری از توالی ردیابی شده به طور متوسط قبل از از دست دادن هدف)، Acc.[12] (همپوشانی متوسط بین پیش‌بینی رهگیر و مقدار واقعی در فریم‌هایی که با موفقیت ردیابی شده‌اند) و EAO[13] (هم‌پوشانی متوسط مورد انتظار) که از مهم‌ترین معیارهای ارزیابی روش‌های مختلف است اندازه‌گیری می‌شود [۶۳]. همان‌طور که در جدول ۱ تا جدول ۶ نشان داده شده، رویکرد STD-Siam با چندین رهگیر مطرح فعلی در پایگاه‌داده‌های مختلف، از جمله SFTransT [۵۳]، SiamTC [۵۲]، SiamRTU [۵۱]، STMTrack [۳۰]، D3S [۴۶]، SiamMask [۴۹]، ATOM [۴۲]، SiamRPN++ [۴۱]، SiamBAN [۶۴]، SiamFC [۱۴]، SiamTPN [۶۵]، LightTrack [۳۱] ، SiamFC++ [۶۶] و SiamGAT [۶۷] در چالش‌های مختلف مقایسه، و روش پیشنهادی در اکثر نتایج بهترین امتیاز را به‌دست آورده است.

جدول ۱: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر پیشنهادی (STD-Siam) در پایگاه داده VOT2019/2020 (دو رهگیر با بالاترین امتیاز به صورت برجسته مشخص شده اند.)

	EAO ↑	Acc. ↑	Rob. ↓	FPS[14]
STD-Siam	**0.502**	**0.714**	**0.358**	66
D3S	**0.441**	**0.698**	0.765	25
SiamTC	0.345	0.594	**0.371**	72
LightTrack	0.333	0.536	0.321	38
SiamMask	0.321	0.625	0.645	55
SiamFC++	0.288	0.583	0.406	**160**
DiMP	0.314	0.582	0.371	40
ATOM	0.271	0.462	0.734	30
SiamRPN++	0.285	0.598	0.482	35
SiamBAN	0.327	0.601	0.396	40
SiamFC	0.179	0.501	0.419	**86**

همچنین رهگیر ارائه شده با پایگاه داده‌های VOT2016 [۳۲] و VOT2018 [۳۳] آزمایش شده است. هر مجموعه داده شامل ۶۰ دنباله چالش برانگیز است. در این روش ارزیابی، رهگیر می‌تواند در صورت عدم موفقیت مجدداً راه‌اندازی شود تا پس از راه‌اندازی مجدد آنالیز شود. جابه‌جایی بین دو روش سنتی و یادگیری ژرف باعث شده است که عملکرد رهگیر STD-Siam بهتر از روش‌های دیگر در تمام معیارهای VOT مورد بررسی قرار گیرد. بدون شک یادگیری برخط و برافزایی داده‌ها نقش موثری در این موفقیت داشته است. نتایج در جدول ۲ و جدول ۳ نشان داده شده است. مقایسه با رهگیر اصلی SiamFC برتری روش پیشنهادی را نشان می‌دهد. SiamFC برای رویارویی با تغییرات ظاهری هدف در بلندمدت نیاز به یادگیری برخط دارد که در روش اصلی این امکان وجود ندارد. تکنیک‌های تولید داده آموزشی به روش پیشنهادی کمک کرده تا نسبت به بسیاری از رهگیرها به صورت برخط و بلندمدت عملکرد بهتری داشته باشد.

جدول ۲: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر پیشنهادی در پایگاه داده VOT2018.

	EAO ↑	Acc. ↑	Rob. ↓	FPS
STD-Siam	**0.532**	**0.708**	**0.131**	66
D3S	**0.488**	**0.641**	**0.153**	25
STMTrack	0.447	0.590	0.159	37
SiamRTU	0.423	0.603	0.215	20
SiamRPN++	0.413	0.601	0.234	35
ATOM	0.400	0.590	0.205	30
SiamFC++	0.401	0.556	0.183	**160**
SiamMask	0.381	0.610	0.277	55
SiamRPN	0.245	0.439	0.461	**200**
SiamFC	0.188	0.509	0.585	86

جدول ۳: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر پیشنهادی در پایگاه داده VOT2016.

	EAO ↑	Acc. ↑	Rob. ↓	FPS
STD-Siam	**0.529**	**0.716**	**0.124**	66
D3S	**0.492**	**0.660**	**0.132**	25
SiamMask	0.430	0.641	0.214	55
ATOM	0.431	0.611	0.179	30
SiamRPN++	0.370	0.581	0.240	35
SiamRPN	0.343	0.560	0.303	**200**
SiamFC	0.235	0.530	0.381	**86**

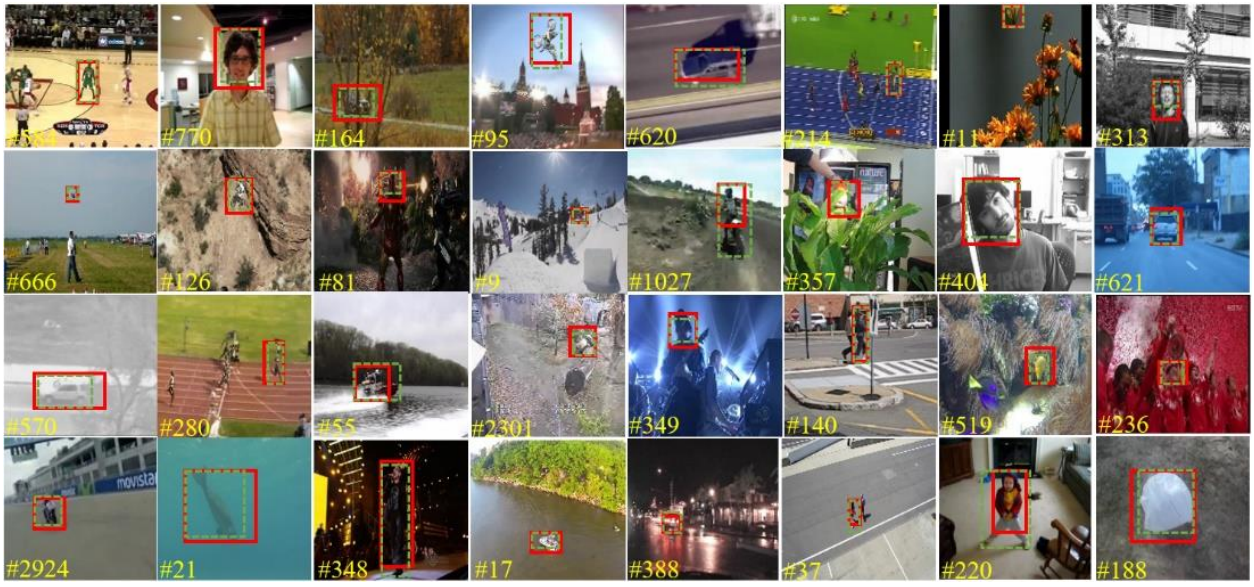
پایگاه داده رهگیری هدف (OTB-100) [۲۴] یک معیار استاندارد برای ردیابی بصری است که شامل ۱۰۰ هدف با ۱۱ ویژگی مختلف است. همان‌طور که در جدول ۴ نشان داده شده است، روش پیشنهادی به AUC[15] 72.3% دست یافته که نزدیک به SiamRPN++ و همچنین بهتر از بسیاری از الگوریتم‌های مرسوم است.

جدول ۴: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر ارائه شده در این مقاله در پایگاه داده OTB100.

	AUC (%)	FPS
STD-Siam	**72.3**	66
SFTransT	**70.3**	27
SiamRPN++	69.6	35
TransT	69.3	65
MDNet	67.9	55
ATOM	66.1	30
SASiam	65.8	50
SiamRPN	63.6	**200**
SiamFC	58.3	**86**
CFNet	56.8	75

پایگاه داده LaSOT [۶۸] یک مجموعه داده با کیفیت بالا و در مقیاس بزرگ است که شامل ۱۴۰۰ چالش با توالی طولانی است. این مجموعه داده دارای ۱۱۲۰ داده آموزشی و ۲۸۰ چالش برای آزمایش است. عملکرد رهگیرهای

[11] Robustness or Failure Rate
[12] Accuracy
[13] Expected Average Overlap
[14] Frame Per Second
[15] Area Under Curve



شکل ۵: نتایج کیفی رهگیری با استفاده از روش ارائه شده روی دنباله‌های Bolt1، Bolt2، Skiing، Da-vid، Carscale، بسکتبال و... . رهگیر پیشنهادی می‌تواند به‌طور موثر هدف را در این دنباله‌های پر چالش تعقیب کند. کادر قرمز نشان دهنده جعبه محاطی STD-Siam است و کادر سبز جعبه مرزی داده‌های مرجع درستی است.

مختلف با استفاده از معیار موفقیت (AUC) ، دقت و امتیاز دقت بهنجارسازی شده در جدول ۵ نشان داده شده است. کسب بالاترین امتیاز AUC 81.7 درصد توسط روش رهگیری پیشنهادی در این معیار ردیابی با سرعت بالا و طولانی‌مدت، ناشی از آموزش برخط شبکه در طی رهگیری دارد. البته برافزایی داده‌ها و کمک به آموزش شبکه را نمی‌توان در دستیابی به این موفقیت نادیده گرفت.

جدول ۵: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر ارائه شده در این مقاله در پایگاه داده LaSOT.

	Precision(%)	Norm. Prec.(%)	Success (AUC) (%)	FPS
STD-Siam	**78.9**	**89.3**	**81.7**	66
SiamRPN++	69.4	**80.1**	**73.2**	35
MixFormer	**76.3**	79.9	70.1	18
SFTransT	73.9	78.1	69.0	27
KeepTrack	70.2	77.2	67.1	18
TransT	69.01	73.7	64.9	50
DaSiamRPN	59.0	73.2	63.7	**110**
MDNet	56.4	70.5	60.6	1
SiamGAT	53.9	63.3	53.0	70
SiamFC	53.3	66.2	57.0	**86**
CFNet	53.2	65.4	57.9	83
DiMP	56.6	65.0	57.0	40

پایگاه داده GOT-10k [۳۴] یک مجموعه داده در مقیاس بزرگ و با تنوع بالا است که از ۱۰ هزار توالی ویدیویی با اهداف مشخص شده توسط جعبه‌های محاطی تشکیل شده است. رهگیرها بر روی ۱۸۰ دنباله آزمایشی با ۸۴ کلاس شی مختلف و ۳۲ الگوی حرکت ارزیابی می‌شوند، در حالی‌که فقط در فریم اول هر رهگیر باید مقداردهی اولیه شود. در مجموعه داده GOT-10k ، روش پیشنهادی تقریباً فاصله کمی با ردیاب TrDiMP دارد که نتیجه قابل قبولی است. نتایج مقایسه در جدول ۶ نشان داده شده است و STD-Siam توانسته موفق به کسب AO 66.2% در GOT-10k شود.

جدول ۶: مقایسه عملکرد رهگیرهای مشابه فعلی با رهگیر ارائه شده در این مقاله در پایگاه داده GOT-10k.

	AO[17] ↑	$SR_{0.75}$[16]↑	$SR_{0.5}$ ↑	FPS
SFTransT	**72.7**	**66.9**	**84.3**	27
TrDiMP	**67.1**	**58.3**	**77.7**	26
STD-Siam	66.2	57.4	76.5	66
STMTrack	64.2	57.5	73.7	37
SiamTC	63.5	51.5	74.3	72
SiamFC++	59.5	47.9	69.5	90
D3S	59.7	46.2	67.6	25
ATOM	55.6	40.2	63.5	30
SiamMask	51.4	36.6	58.7	55
SiamRPN	46.3	25.3	54.9	**200**
SiamFC	34.8	9.8	35.3	86
GOTURN	34.3	12.3	37.6	**100**
MDNET	29.8	10.0	30.3	1

نتایج کیفی رهگیری شی برای برخی از سکانس‌ها توسط STD-Siam در شکل ۵ نشان داده شده است. STD-Siam می‌تواند دائماً هدف دلخواه را با موفقیت رهگیری کند و به‌طور کلی عملکرد بهتری نسبت به سایر روش‌ها دارد زیرا رهگیر سنتی چندین فریم را که شی در آن قرار دارد ردیابی و به واحد تولید داده‌های برافزوده ارسال می‌کند. سپس با افزایش داده‌ها، واحد آموزش شروع به‌کار می‌کند. اگر تعداد فریم‌ها خیلی کم باشد، ردیاب یادگیری ژرف فعال نمی‌شود. نتایج ثابت می‌کند که رهگیر STD-Siam می‌تواند بسیاری از اهداف غیر صلب را در محیط‌های پیچیده ردیابی کند.

۵-نتیجه‌گیری

در این مقاله روش STD-Siam برای رهگیری بصری شی به‌طور طولانی مدت و بی‌درنگ معرفی شده که به‌صورت برخط آموزش می‌بینند. عملیات رهگیری با راه‌گزین بین دو روش یادگیری عمیق و روش سنتی انجام می‌گیرد. در فرآیند رهگیری، هدف مورد نظر با استفاده از روش دسته‌بندی حداقل مربعات منظم شده رهگیری می‌شود و نتایج داده‌های برچسب زده شده به واحد برافزایی داده‌ها ارسال می‌شود. نقص داده‌های ناکافی جهت آموزش با استفاده از تولید داده‌های آموزشی حل می‌شود و با برش تصاویر می‌توان به طور قابل

[16] Success Rates (SR) at the Overlap Thresholds 0.75

[17] Average Overlap

توجهی بر چالش انسداد نسبی هدف در رهگیری غلبه کرد. پس از اتمام یادگیری، رهگیر یادگیری ژرف روند ردیابی هدف را ادامه می‌دهد. علاوه بر این، روبش کل تصویر با آشکارساز نقشه برجستگی، از تمام اطلاعات برای شناسایی سریع هدف در تصویر استفاده می‌کند، بدون اینکه نیاز به اجرای پنجره کشویی جهت روبش تصویر باشد. STD-Siam توانسته با سرعت ۶۶ فریم‌درثانیه اجرا شود و در مقایسه با سایر رهگیرهای مشابه فعلی، نتایج قابل قبولی در پایگاه داده‌های بزرگ مانند VOT2019/2020 ، VOT2018، VOT2016، OTB100، LaSOT و GOT-10K به‌دست آورد. در آینده قصد داریم با استفاده از ادغام داده‌های حسگر راداری و دوربین، دقت و قدرت رهگیری را افزایش دهیم تا عملکرد رهگیر در شرایط نامساعد جوی بهبود یابد.

مراجع

[1] Liu, L., Xing, J., and Ai, H.: Multi-view vehicle detection and tracking in crossroads. in *Proceedings of the Asian Conference on Pattern Recognition* (ACPR), pp. 608–612 (2011)

[2] Liu, L., Xing, J., Ai, H., and Ruan X.: Hand posture recognition using finger geometric feature. in *IEEE International Conference on Pattern Recognition* (ICPR), pp. 565–568 (2012)

[3] Emami, A., Dadgostar, F., Bigdeli, A., and Lovell, B.: Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance. in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pp. 349–354 (2012)

[4] Zhang, m., Xing, J., Gao, J., and Hu, W.: Robust visual tracking using joint scale-spatial correlation filters. in *IEEE International Conference on Image Processing (ICIP)* (2015)

[۵] وحید آزادزاده، علی محمد لطیف، «دسته‌بندی ویژگی‌های استخراج شده از پیش‌زمینه و پس‌زمینه تصویر برای ردیابی اهداف متحرک هوایی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۳، پاییز ۱۳۹۵

[۶] عقیل عبیری، محمدرضا محزون، « ردیابی اهداف متحرک هوایی با استفاده از تخمین چگالی کرنل بر اساس الگوریتم فیلتر ذره»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، شماره ۳، پاییز ۱۳۹۴.

[7] Guo, Q., Feng, W., Zhou, C., Pun, C.M., and Wu, B.: Structure-regularized compressive tracking with online data-driven sampling. in *IEEE Transactions on Image Processing*, pp. 5692-5705 (2017)

[8] Zhang, T., Liu, S., Ahuja, N., Yang, M.H. and Ghanem, B.: Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision*, 111(2) pp. 171–190 (2015)

[9] Wang, N., Shi, J., Yeung, D.Y., and Jia, J.: Understanding and diagnosing visual tracking systems, in *International Conference on Computer Vision (ICCV*) (2015)

[10] Kristan, M., Matas, J., Leonardis, A., *et al.*: The visual object tracking VOT2015 challenge results. in *International Conference on Computer Vision Workshops (ICCVW) (2015)*

[11] Wang, L., Ouyang, W., Wang, X., and Lu, H.: STCT: Sequentially training convolutional networks for visual tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2016)

[12] Ma, C., Huang, J.B., Yang, X., and Yang, M.H.: Hierarchical convolutional features for visual tracking. in *International Conference on Computer Vision (ICCV*) (2015)

[13] Hong, S., You, T., Kwak, S., and Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. in *International Conference on Machine Learning* (ICML) (2015)

[14] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., and Torr, P.H.: Fully-convolutional Siamese networks for object tracking. in *European Conference on Computer Vision* (ECCV Workshops) (2016)

[15] Cui, Y., Jiang, C., Wang, L., and Wu, G.: MixFormer: End-to-end tracking with iterative mixed attention. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn tion (CVPR)* (2022)

[16] Guo, D., Wang, J., Cui, Y., Wang, Z., and Chen, S.: SiamCAR: Siamese fully convolutional classification and regression for visual tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2020)

[17] Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., and Hu, W.: Distractor-aware Siamese networks for visual object tracking. in *European Conference on Computer Vision* (2018)

[18] Nam, H., and Han, B.: Learning multi-domain convolutional neural networks for visual tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2016)

[19] Danelljan, M., Häger, G., Khan, F.S., and Felsberg, M.: Convolutional features for correlation filter based visual tracking. in *International Conference on Computer Vision Workshops (ICCVW) (2015)*

[20] Danelljan, M., Hager, G., Khan, F., Felsberg, M.: Convolutional features for correlation filter based visual tracking. in: ICCV 2015 Workshop, pp. 58–66 (2015)

[21] Chen, K., and Tao, W.: Once for all: A two-flow convolutional neural network for visual tracking. in *arxiv:1604.07507* (2016)

[22] Wang, G., Luo, C., Xiong, Z., and Zeng, W.: Spm-tracker: Series-parallel matching for real-time visual object tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2019)

[23] Danelljan, M., Bhat, G., Shahbaz Khan, F., and Felsberg, M.: ECO: Efficient convolution operators for tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2017)

[24] Wu, Y., Lim, J., and Yang, M.H.: Object tracking benchmark. in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(9), pp. 1834–1848 (2015)

[25] Zhu, H., Xue, M., *et al.*: Fast Visual Tracking with Siamese Oriented Region Proposal Network. in *IEEE Signal Processing Letters*, pp. 1437–1441 (2022)

[26] Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., and Maybank, S.: Learning attentions: Residual attentional Siamese network for high performance online visual tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2018)

[27] Wang, Q., Zhang, M., Xing, J., Gao, J., Hu, W., and Maybank, S.: Do not lose the details: Reinforced representation learning for high performance visual tracking. in *International Joint Conference on Artificial Intelligence* (IJCAI) (2018)

[28] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. in *Conference and Workshop on Neural Information Processing Systems* (NIPS) (2012)

[29] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2016)

on Computer Vision and Pattern Recognition (*CVPR)*, pp. 6578–6588 (2020)

[48] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., and Yan, J.: SiamRPN++: Evolution of Siamese visual tracking with very deep networks. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)*, pp. 4282–4291 (2019)

[49] Wang, Q., Zhang, L., Bertinetto L., *et al.*: Fast online object tracking and segmentation: A unifying approach. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2019)

[50] Goutam, B., Danelljan, M *et al.*: Learning discriminative model prediction for tracking. in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6182–6191 (2019)

[51] Zhao, F., *et al.*: Siamese Regression Tracking with Reinforced Template Updating. in *IEEE Transactions on Image Processing*, pp. 628–640 (2020)

[52] Jiang, Y., Song, X., *et al.*: Target-Cognisant Siamese Network for Robust Visual Object Tracking. *Pattern Recognition Letters*, vol. 163, pp. 129-135 (2022)

[53] Tang, C., et al. Learning spatial-frequency transformer for visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology* (2023)

[54] Wang, N., Zhou, W., Wang, J., and Li, H.: Transformer meets tracker: Exploiting temporal context for robust visual tracking. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (*CVPR)*, pp. 1571–1580 (2021)

[55] Cui, Y., Jiang, *et al.*: End-to-end tracking with iterative mixed attention. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13608–13618, (2022)

[56] Chen, X., Yan, B., Zhu, J., Wang, D., *et al*.: Transformer tracking. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (*CVPR)*, pp. 8126–8135 (2021)

[57] Mayer, C., Danelljan, M., Paudel, D.P., and Gool, L.V.: Learning target candidate association to keep track of what not to track. in *Proceedings of the IEEE International Conference on Computer Vision*, (ICCV), pp. 13444–13454 (2021)

[58] Chen, Y., Wang, C.Y., Yang, C.Y., *et al.*: NeighborTrack: Improving Single Object Tracking by Bipartite Matching with Neighbor Tracklets. in *arXiv:2211.06663*, (2022)

[59] Hou, X., Lim, J., and Zhang, L.: Saliency detection: A spectral residual approach. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2007)

[60] Chen, P., et al. "Gridmask data augmentation." *arXiv preprint arXiv:2001.04086* (2020)

[61] Kristan, M., Matas, J., Leonardis, A., *et al.*: The seventh visual object tracking vot2019 challenge results. in *International Conference on Computer Vision Workshops (ICCVW) (2019)*

[62] Kristan, M., Leonardis, A., *et al*.: The eighth visual object tracking VOT2020 challenge results. in *European Conference on Computer Vision* (ECCV) (2020)

[63] Javed, S., Danelljan, M., *et al.*: Object Tracking With Discriminative Filters and Siamese Networks: A Survey and Outlook. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6552-6574, doi: 10.1109/ TPAMI. 2022.3212594 (2023)

[64] Chen, Z., Zhong, B., Li, G., Zhang, S., and Ji, R.: Siamese box adaptive network for visual tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2020)

[30] Fu, Z., Liu, Q., *et al.*: Stmtrack: Template-free visual tracking with space-time memory networks. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13774–13783 (2021)

[31] Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H.: Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15189 (2021)

[32] Kristan, M., Leonardis, A., Matas, J., *et al.*: The visual object tracking VOT2016 challenge results. in *European Conference on Computer Vision* (ECCV) (2016)

[33] Kristan, M., Leonardis, A., *et al.*: The sixth visual object tracking VOT2018 challenge results. in *European Conference on Computer Vision* (ECCV), pp. 3–53 (2018)

[34] Huang, L., Zhao, X., and Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019)

[35] Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. in *Computer Vision and Pattern Recognition* (*CVPR)* (2008)

[36] Fan, J., Xu, W., Wu, Y., Gong, Y.: Human tracking using convolutional neural networks. in *IEEE Transactions on Neural Networks* 21(10), pp.1610-1623 (2010)

[37] Bagherzadeh, M.A., Yazdi, M.: Fast object tracking with long-term occlusions handling in dynamic scenes. in *International Conference on Robotics and Mechatronics (ICRoM)* (2014)

[38] Bolme, D.S., Beveridge, J.R., Draper, B.A., and Lui, Y.M.: Visual object tracking using adaptive correlation filters. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2010)

[39] Henriques, J.F., Caseiro, R., Martins, P., and Batista, J.: High-speed tracking with kernelized correlation filters. in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(3), pp. 583–596 (2015)

[40] Bagherzadeh, M.A., and Yazdi, M.: Regularized least-square object tracking based on ℓ2, 1 minimization. *in IEEE International Conference on Robotics and Mechatronics (ICROM)* (2015)

[41] Hu, W., Wang, Q., *et al.*: DCFNet: Discriminant correlation filters network for visual tracking. in *Journal of Computer Science and Technology*, Doi :10.1007/s11390-023-3788-3 (2023)

[42] Danelljan, M., Bhat, G., Khan, F.S., and Felsberg, M.: Atom: Accurate tracking by overlap maximization. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2019)

[43] Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese Instance Search for Tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2016)

[44] Held, D., Thrun, S., and Savarese, S.: Learning to track at 100 fps with deep regression networks. in *European Conference on Computer Vision* (ECCV) (2016)

[45] Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X.: High performance visual tracking with Siamese region proposal network. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2018)

[46] Lukezic, A., Matas, J., and Kristan, M.: D3S-a discriminative single shot segmentation tracker. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2020)

[47] Voigtlaender, P., Luiten, J., Torr, P.H., and Leibe, B.: Siam R-CNN: Visual tracking by re-detection. in *IEEE Conference*

[65] Xing, D., Evangeliou, N., Tsoukalas, A., Tzes, A.: Siamese transformer pyramid networks for real-time UAV tracking. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2139–2148 (2022)

[66] Xu, Y., Wang, Z., Li, Z., Ye, Y., and Yu, G.: Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines: in *AAAI Conference on Artificial Intelligence* (2020)

[67] Guo, D., Shao, Y., Cui, Y., *et al.*: Graph attention tracking. in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)* (2021)

[68] Fan, H., Lin, L., Yang, F., *et al.*: A high-quality benchmark for large-scale single object tracking. in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)* (2019)