DOR: 20.1001.1.2538161.1403.9.4.2.7
DOI : https://doi.org/10.22034/csj.2025.214933
زمستان ۱۴۰۳، دوره ۹، شماره ۴، پیاپی ۳۵

دریافت مقاله: ۱۴۰۳/۰۹/۲۸
پذیرش مقاله: ۱۴۰۳/۱۱/۰۱

نوع مقاله: پژوهشی

ناوبری خودمختار پهپاد در گذرگاه‌های باریک بدون سامانه موقعیت‌یابی جهانی با یادگیری تقویتی عمیق

مهدی شهبازی خجسته
دانش‌آموخته کارشناسی ارشد دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
پست الکترونیکی: m.shahbazikhojasteh@mail.sbu.ac.ir

آرمین سلیمی بدر*
استادیار دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
پست الکترونیکی: a_salimibadr@sbu.ac.ir

چکیده

هواپیمای بدون سرنشین (پهپاد) به دلیل هزینه‌های عملیاتی پایین و انعطاف‌پذیری بالا، در زمینه‌هایی نظیر نقشه‌برداری، جستجو و نجات، نظارت و اکتشاف کاربرد گسترده‌ای یافته‌اند. با این وجود، ناوبری کارآمد و عبور موفق از گذرگاه‌های باریک در محیط‌های پیچیده، بدون استفاده از حسگرهای گران‌قیمت همچنان یک چالش اساسی محسوب می‌شود. تحقیقات پیشین اغلب به استفاده از حسگرهای پیشرفته یا روش‌های غیربهینه مبتنی بوده‌اند که عملکرد ضعیفی در شرایط متغیر محیطی نشان می‌دهند. این رویکردها برای جلوگیری از تشخیص نادرست موانع و بهبود قابلیت اطمینان و ایمنی بر چندین حسگر مانند فراصوت، مادون‌قرمز، لیدار و دوربین تکیه می‌کنند. این روش‌ها به دلیل نیاز به پردازش حجم زیادی از داده‌ها، دارای سربار محاسباتی بالا هستند و عملکرد آن‌ها در شرایط محدودیت منابع بهینه نیست. علاوه بر این، این روش‌ها معمولاً از قابلیت تعمیم‌پذیری کافی برخوردار نیستند. از طرف دیگر، رویکردهای برنامه‌ریزی مسیر نظیر روش‌های مبتنی بر جستجو در گراف، میدان پتانسیل و مبتنی بر نمونه‌برداری، علی‌رغم تحلیل‌پذیری، در تطبیق با تغییرات مدل یا محیط دچار چالش می‌شوند و به مکان‌های مقیاس کوچک محدود می‌شوند و برای کاربردهای دنیای واقعی دقت کافی ندارند. به همین منظور، در این مقاله یک سامانه ناوبری خودمختار مبتنی بر یادگیری تقویتی عمیق و استفاده از تصاویر عمقی ارایه شده است. در این روش، مسئله تصمیم‌گیری به‌عنوان یک فرآیند تصمیم‌گیری مارکوف مدل‌سازی می‌شود تا پهپاد با اتکا به اطلاعات کنونی قابلیت تصمیم‌گیری را داشته باشد. همچنین، از شبکه‌های عصبی هم‌آمیختی برای استخراج ویژگی‌های بصری از تصاویر عمقی استفاده می‌شود. برای اتخاذ یک سیاست تصمیم‌گیری مناسب، از الگوریتم بهینه‌سازی سیاست تقریبی برای آموزش مدل بهره گرفته شده است. سامانه پیشنهادی در یک محیط شبیه‌سازی‌شده واقع‌گرایانه آزمایش شده و عملکرد آن مورد ارزیابی قرار گرفته است.

* نویسندهٔ مسئول

نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی توانسته است با ۸۵ درصد موفقیت و ۱۵ درصد میزان برخورد، مسیر گذرگاه‌های باریک را طی کند و با ۱۷ درصد بهبود، عملکرد بهتری نسبت به روش‌های قیاس شده مرجع ارایه دهد. علاوه بر این، افزایش مداوم پاداش در طول فرآیند آموزش، کاهش بار محاسباتی و قابلیت تعمیم‌پذیری به شرایط محیطی مختلف، نظیر تغییر در پیکربندی محیط یا شکل گذرگاه‌ها، از دیگر مزایای این روش به شمار می‌روند. رویکرد پیشنهادی در آزمایش‌های تعمیم‌پذیری توانسته موفقیت ۱۰۰ درصد را کسب کند. این رویکرد می‌تواند به‌عنوان یک راهکار مقرون‌به‌صرفه و کارآمد به‌ویژه در شرایطی که استفاده از حسگرهای گران‌قیمت امکان‌پذیر نیست مورد استفاده قرار گیرد.

واژه‌های کلیدی: هواپیمای بدون سرنشین، مسیریابی خودکار، اجتناب از مانع، یادگیری تقویتی عمیق، بهینه‌سازی سیاست تقریبی

۱.مقدمه

وسایل نقلیه هوایی بدون سرنشین (پهپادها) به‌عنوان ابزارهای متحول‌کننده در کاربردهای مختلف ظاهر شده‌اند و بر هر دو بخش‌های نظامی و غیرنظامی تاثیر می‌گذارند [۱]. اخیراً تقاضا برای پهپادهای کوچک در محیط‌های داخلی و پیچیده افزایش یافته است. استفاده از پهپادهای کوچک به دلیل چابکی، مانورپذیری پیشرفته، مقرون‌به‌صرفه بودن و اندازه کوچک، انجام وظایف مختلفی مانند جستجو و نجات، نقشه‌برداری محیطی و اکتشاف مناطق ناشناخته را ممکن می‌سازد [۱, ۲].

در حالی‌که پهپادها چشم‌اندازهایی خوبی از خود نشان می‌دهند، عملکرد آن‌ها به دلیل محدودیت‌هایی مانند استقامت، باتری و زمان پرواز کاسته شده است [۳]. برای افزایش قابلیت‌های پهپاد، انتخاب دقیق حسگرهای خارجی ضروری است. پهپادها معمولاً به حسگرهایی مانند شتاب‌سنج، مغناطیس‌سنج، ژیروسکوپ و سامانه موقعیت‌یابی جهانی[1] برای ناوبری متکی هستند [٤]. علاوه بر این، آن‌ها می‌توانند از دوربین‌های دوچشمی، فناوری لیدار[2]، حسگرهای فراصوت و فاصله‌سنج برای جمع‌آوری داده‌های محیطی استفاده کنند [٥].

با وجود کاربرد آن‌ها، بسیاری از این حسگرها دارای معایبی هستند که از جمله آن می‌توان به هزینه بالا، وزن قابل توجه، افزایش مصرف انرژی و پیچیدگی عملیاتی اشاره نمود؛ اما در مقابل دوربین یک جزء حیاتی است [٤, ٦] که یک رویکرد همه‌کاره و کارآمد برای افزایش ادراک محیطی پهپاد ارایه می‌دهد. دوربین‌ها به‌ویژه برای کاربردهایی که به حداقل رساندن انرژی، اندازه و وزن حائز اهمیت است، بسیار مناسب هستند [۷].

پهپادها به‌طور فزاینده‌ای در برنامه‌هایی استفاده می‌شوند که به میزان بالایی از استقلال نیاز دارند. روش‌های سنتی برای ناوبری پهپاد اغلب در محیط‌های پیچیده، به‌ویژه در فضاهای محدود مانند گذرگاه‌های باریک، مشکل دارند. این رویکردها که معمولاً بر مسیرهای پروازی از پیش برنامه‌ریزی شده یا سامانه‌های مبتنی بر قوانین تکیه دارند، محدودیت‌هایی را در سازگاری و پاسخگویی به موانع پیش‌بینی نشده یا تغییرات محیطی نشان می‌دهند [۲, ۸]. برای مدیریت این محدودیت‌ها، ادغام هوش مصنوعی، به‌ویژه یادگیری ماشین و یادگیری عمیق، راه‌های امیدوارکننده‌ای برای افزایش استقلال پهپاد و قابلیت‌های ناوبری ارایه می‌دهد [۵, ۹, ۱۰]. یادگیری ماشین پهپادها را قادر می‌سازد تا از داده‌ها یاد بگیرند و آن‌ها را قادر می‌سازد تا با محیط‌ها سازگار شوند و بدون برنامه‌نویسی صریح تصمیم‌گیری آگاهانه بگیرند [۱۱]. یادگیری عمیق از شبکه‌های عصبی مصنوعی با لایه‌های متعدد برای استخراج الگوها و بازنمایی‌های پیچیده از داده‌ها استفاده می‌کند که منجر به افزایش قابلیت‌هایی خودمختاری می‌شود [۹, ۱۲].

برای اطمینان از این که یک پهپاد می‌تواند وظایف

1- Global Positioning System (GPS)
2- Light Detection and Ranging (LiDAR)

محولـــه را به‌طـــور ایمن و کارآمد انجام دهـــد، باید بتواند بدون کمک انسان تصمیم بگیرد و با شرایط متغیر سازگار شود. در این زمینه، یادگیری تقویتی[3] پایه‌ای برای یادگیری سیاست‌های تصمیم‌گیری بهینه فراهم می‌کند. با ترکیب هر دو روش، یادگیری تقویتی عمیق[4] شکل می‌گیرد که پتانسیل قابل‌توجهی برای دســـتیابی به سطوح بالاتر استقلال ارایه می‌دهد [۲, ۱۳].

یادگیـــری تقویتی عمیق یک عامل را قادر می‌ســـازد تا اقدامات بهینه را از طریق تعامل با محیط و دریافت پاداش در رفتارهای مطلـــوب و تنبیه بـــرای رفتارهای نامطلوب بیاموزد [۱٤]. این فرآیند یادگیری با آزمون‌وخطا به عامل اجازه می‌دهد تا سیاســـت تصمیم‌گیری خـــود را به‌تدریج اصلاح کند و عملکرد ناوبری و اجتناب از مانع را در طول زمان بهبود بخشد.

مطالعـــات متعـــددی رویکردهای یادگیـــری تقویتی را برای کنترل و ناوبری وســـایل نقلیه خودران زمینی [۱۵–۱۷] و پهپـــاد [۱۸–۲۵] اتخاذ کرده‌اند کـــه اثربخشی آن را در تصمیم‌گیری هوشـــمند و ســـازگاری با وضعیت‌های چالش‌برانگیز نشـــان می‌دهد. به‌منظور اســـتفاده کامل از دوربین‌هـــای تک‌چشـــمی، الگوریتم‌های اجتنـــاب از مانع مختلفی توسعه داده شده است. با این حال، استفاده مستقیم از اطلاعات رنگی از یک دوربین تک‌چشـــمی چالش‌برانگیز اســـت. زیرا، دوربین‌های تک‌چشمی داده‌های دنیای واقعی ســـه‌بعدی را به پیکسل‌های تصویر دوبعدی تبدیل می‌کنند که دریافت اطلاعات فاصله دقیق از موانع را دشوار می‌کند

بـــرای مقابله با این مشـــکل، مطالعات مختلفی شـــکل گرفته اســـت. به‌عنوان نمونه، [۲٦] با اســـتفاده از تصاویر رنگی دریافتی از یک دوربین تک‌چشـــمی، نقشه‌های عمق متراکم می‌سازند و بر اساس نقشه‌های عمق ساخته شده، برنامه‌ریـــزی مسیر بدون مانع تولید می‌شـــوند. پهپاد در بیشتر موارد با موفقیت از موانع اجتناب می‌کند. با این حال، مدت زمان مورد نیاز برای محاسبه این مسیرها زیاد است که می‌تواند منجر به شکست برای محیط‌های پیچیده شود. مطالعات [۲۷] و [۲۸] از تکنیک شـــار بصری[5] برای اجتناب از مانع استفاده کرده‌اند. این رویکرد رفتار حشرات را تقلید می‌کنـــد. از آنجایی که این روش بار محاســـباتی کمی نیاز دارد، برای پهپادهای کوچک که سخت‌افزارهای محاسباتی با عملکرد قوی ندارند، مناسب است. با این حال، اعمال شار بصری در محیط‌های بدون بافت یا موانع بزرگ نامناسب اســـت. زیرا نقاط ویژگی در این شرایط به‌راحتی استخراج نمی‌شوند [۲۹].

بـــا وجود تلاش‌هـــای انجام شـــده، روش‌های موجود همچنان چالش‌های متعددی دارند. برای مثال، محاســـبات ســـنگین موردنیاز برخی روش‌ها، مانند ساخت نقشه‌های عمق متراکم زمـــان پاســـخگویی را در محیط‌های پیچیده کاهـــش می‌دهد. همچنیـــن، کاربرد رویکردی نظیر شـــار بصـــری محدودتر و مبتـــنی بر برقراری شـــرایط خاص محیطی است. این محدودیت‌ها ضرورت ارایه رویکردی را نشان می‌دهد که بتواند با بهره‌گیری از حداقل سخت‌افزار و محاسبات، و بخصوص حسگرهای کمتر، ناوبری ایمن و سریع را در محیط‌های دشوار تضمین کند.

بـــا توجه به این چالش‌ها و نقاط ضعـــف، این مقاله به مسئله ناوبری خودکار پهپاد در گذرگاه‌های باریک در یک محیط فاقد سامانه موقعیت‌یابی جهانی با استفاده از رویکرد یادگیری تقویتی عمیق مبتنی بر بینایی می‌پردازد. علاوه بر این، عملکرد یادگیری تقویـــتی عمیق در تصمیم‌گیری‌های پهپاد، با تمرکز بر معیارهای کلیدی مانند میزان موفقیت و اجتناب از برخورد، بررسی می‌شود. مشارکت‌های اصلی این مقاله به شرح زیر است:

- یک ســـامانه ناوبری مبتنی بر یادگیـــری تقویتی عمیق پیاده‌ســـازی شده اســـت که به جهت اجتناب از موانع تنها از حسگر دوربین برای استخراج اطلاعات فضایی تصاویر اســـتفاده می‌کند. در رویکردهای پیشـــین بـــرای اجتناب از موانع از چندین حســـگر مختلف نظیـــر لیدار و دوربین اســـتفاده شده اســـت. در این مقاله برخلاف این رویکردها

3-Reinforcement Learning (RL)
4-Deep Reinforcement Learning (DRL)
5-Optical Flow

تنها از دوربین تک‌چشمی برای انجام ناوبری استفاده شده است. این انتخاب به دلیل کاهش وزن، هزینه و مصرف انرژی، به‌ویژه برای پهپادهای کوچک و با منابع محدود ضروری است. نتایج نشان‌دهنده ناوبری قابل‌اعتماد رویکرد پیشنهادی است. همچنین، قابلیت تعمیم‌پذیری و سازگاری این سامانه از طریق آزمایش‌ها رؤیت شده است؛

- به‌منظور آسان‌سازی و تسریع در حل مسئله، یک تابع پاداش با تمرکز بر عبور از گذرگاه‌های باریک ارایه می‌شود که با نزدیکی بیشتر به نقطه مرکزی، سیگنال پاداش بزرگ‌تری ارایه می‌کند. در مقایسه با رویکردهای مشابه که رسیدن به مقصد را به‌عنوان پاداش در نظر می‌گیرند، تابع پاداش پیشنهادی همانند یک مربی با ارایه بازخورد نسبت به رفتار عامل در عبور از گذرگاه‌ها، موجب بهبود عملکرد و هدایت ایمن‌تر پهپاد می‌شود.
- مسئله با فرآیند تصمیم‌گیری مارکوف مدل‌سازی می‌شود. این بدان معناست که تاریخچه مسیر طی شده توسط پهپاد تا زمانی که وضعیت فعلی آن مشخص باشد برای تصمیم‌گیری‌های آتی بی‌اهمیت است. علاوه بر آن، از یک معماری سبک‌وزن برای پردازش تصاویر ورودی استفاده می‌شود. در قیاس با تحقیقات پیشین، در رویکرد کنونی استنتاج تنها از یک تک تصویر با ابعاد ۵۰×۵۰ انجام می‌شود که با کاهش بار محاسباتی و ساده‌سازی معماری شبکه عصبی، امکان پردازش تصاویر در زمان واقعی را فراهم می‌کند؛
- در این تحقیق، برخلاف پژوهش‌های پیشین که شبیه‌سازی‌ها در محیط‌های باز و با دسترسی به سامانه موقعیت‌یابی جهانی انجام می‌شدند، تمرکز بر ناوبری در محیط‌های داخلی و بدون بهره‌گیری از این فناوری است که چالش‌های جدیدی را به همراه دارد. با استفاده از شبیه‌سازی سه‌بعدی، واقع‌گرایی مسئله افزایش یافته و شرایط شبیه‌سازی به وضعیت‌های واقعی نزدیک‌تر می‌شود.

ساختار این مقاله بدین شرح است: پیشینه در بخش بعدی بررسی می‌شود. در بخش روش پیشنهادی، روش‌شناسی تشریح می‌شود. نتایج و آزمایش‌ها در بخش نتایج تجربی ارایه و تحلیل می‌شود. در نهایت، بخش جمع‌بندی نکات نهایی را ارایه می‌کند و جهت‌های بالقوه را برای تحقیقات آینده پیشنهاد می‌کند.

۲. کارهای مرتبط

روش‌های سنتی ناوبری پهپاد و اجتناب از موانع با روش‌های برجسته از جمله الگوریتم‌های مبتنی بر گراف، روش‌های میدان پتانسیل و سامانه‌های مبتنی بر قانون به‌طور گسترده مورد مطالعه قرار گرفته‌اند. الگوریتم‌های مبتنی بر گراف مانند الگوریتم‌های دایکسترا و A*، محیط را به یک گراف گسسته تبدیل می‌کنند و از الگوریتم‌های جستجو برای یافتن مسیر بهینه استفاده می‌کنند [۹, ۱۳].

این روش‌ها به دلیل توانایی‌شان در یافتن کوتاه‌ترین مسیر شناخته شده‌اند، اما می‌توانند از ناکارآمدی محاسباتی، به‌ویژه در محیط‌های بزرگ و پیچیده رنج ببرند [۹, ۳۰]. از سوی دیگر، روش‌های میدان پتانسیل [۳۱] روبات را به‌عنوان یک جرم نقطه‌ای در حال حرکت در میدان پتانسیل تولید شده توسط موانع و هدف در نظر می‌گیرند. این روش‌ها از نظر محاسباتی کارآمد هستند اما مستعد حداقل‌های محلی به‌ویژه در زمان حضور موانع مقعر هستند [۸]. سامانه‌های مبتنی بر قوانین برای هدایت حرکت پهپاد بر مجموعه‌ای از قوانین از پیش تعریف شده تکیه می‌کنند. این سامانه‌ها برای پیاده‌سازی ساده هستند اما فاقد انعطاف‌پذیری مناسب هستند [۳۰].

این روش‌ها برای مسیریابی در یک محیط شناخته شده سراسری مناسب‌تر هستند. به‌عنوان نمونه، رویکرد [۳۲] از A* برای یافتن بهترین مسیر برای تحویل غذا مابین تأمین‌کننده و مکان مشتری با پهپاد استفاده می‌کند. اگرچه یک سامانه نمونه اولیه را برای این کار توسعه داد، اما به دلیل نگرانی‌های ایمنی نظیر احتمال برخورد، شکست و خرابی پهپاد در طی انجام وظیفه محوله، از انجام پرواز

خودمختار خودداری کـرد. در عوض، پهپاد را به‌صورت دستی در مسیر بهینه تعیین‌شـده توسط الگوریتم هدایت کرد. الگوریتم A* به‌شـدت به تابع هزینه اکتشافی بستگی دارد و به محاسبات و ذخیره‌سازی مداوم نیاز دارد [۳۳].

علی‌رغـم محدودیت‌ها، الگوریتم A* در طول سـال‌ها با پیشـرفت‌های زیادی روبرو بوده است. به‌عنوان مثال، رویکرد [۳٤] یک مدل برنامه‌ریزی چند هدفه برای ناوبری پهپـاد بـا ترکیب تصحیـح خطـا و محدودیت‌های مسیر پیشنهاد کرد. این الگوریتم در یک محیط نقشه مشبک موثر اسـت اما اسـتفاده از آن در فضاهای با ابعاد بالا کاهش می‌یابـد [۳۵]. بدین‌منظـور، الگوریتم D* بـرای مقابله با محدودیت‌ها معرفی گردید [۳۵]. با این وجود، ایجاد نقشه هزینه برای الگوریتم D* نیازمند پردازش بیشتر و مستعد خطا است [۳۳].

الگوریتم‌هـای حشـره، الگوریتم‌هـای ناوبری موفقی هسـتند که ناوبـری را تجزیه می‌کنند تا به سـمت هدف حرکت کننـد و موانع پیـشِ رو را دور بزنند [۳٦]. اگرچه این رویکردها نیازی به دانش قبلی در مورد محیط ندارند، اما مسیر نهایی آن‌ها از مسیر بهینه فاصله دارد و همچنین فرضیـاتی ایده‌آل برای خروج از مانـع وجود دارد که در دنیای حقیقی یا محیط‌های ناشناخته واقع‌بینانه نیست [۳۷, ۳۸].

از دیگر روش‌ها، درخت جستجوی تصادفی سریع است که یک الگوریتم برنامه‌ریزی مسیر مبتنی بر نمونه‌برداری اسـت [۳۰]. این روش به دلیل رویکرد جستجوی تصادفی خود تحت محدودیت‌های حرکتی، شناخته شده است [۳۵]. مطالعـه [۳۹] آن را با هرس برای حـذف گره‌های اضافی تقویت می‌کند و با محـدود کردن نمونه‌برداری در مناطق امکان‌پذیر، مسیرهای بـدون برخورد را در ناوبری پهپاد مشارکتی امکان‌پذیر می‌دارد. با این حال، این رویکرد فاقد بهینگی مسیر است و در محیط‌های به‌هم‌ریخته یا باریک با چالش روبه‌رو است.

بیشتر رویکردهای سنتی می‌توانند مسیرهای بهینه را پیدا کنند، امـا معمولاً نیاز به دانش کامل از محیط دارند و آن‌ها را برای محیط‌های ناشناخته مناسب نمی‌سازد [٤۰]. در مقابل، رویکردهای یادگیری عمیق مدل آزاد هسـتند و هنگام مواجهه با شرایط جدید یا در حال تغییر، تعمیم قوی را نشان می‌دهند [٤۱]. اگرچه الگوریتم‌های یادگیری عمیق به مجموعه داده‌هـای بزرگ و زمان آموزشی زیادی نیاز دارند، اما پس از آموزش استنتاج سریعی را ارایه می‌دهند. توانایی یادگیری عمیق بـرای یادگیری بازنمایی داده‌های پیچیده از محیط‌های دنیای واقعی، آن را برای کاربردهای روباتیک خودمختار، به‌ویژه برای پهپادها مناسـب می‌کند [۱۲, ٤۲].

شبکه‌های عصبی هم‌آمیختی[6] در استخراج ویژگی‌های فضایی از تصاویر عالی هسـتند و آن‌ها را برای کارهایی مانند تشـخیص موانع و نقشـه‌برداری محیـط که در آن داده‌های بصری بسیار مهم است بسیار موثر می‌سازد. از سوی دیگر، شبکه‌های عصبی بازگشتی[7] داده‌های متوالی را مدیریت می‌کنند و آن‌ها را قادر می‌سازد تا الگوهای زمانی و وابسـتگی‌ها را پردازش کنند که به‌ویژه برای پیش‌بینی مسیر و برنامه‌ریزی حرکت مفید اسـت. برای مثال، [٤۳] یک برنامه‌ریز مسیر سراسری سـریع را پیشنهاد می‌کند که از قابلیت این شبکه برای ایجاد مسیرهای ایمن و بدون برخورد استفاده می‌کند. در مقایسه با الگوریتم‌های سنتی، این رویکرد در محیط‌های چالش‌برانگیز بهتر عمل می‌کند.

مدل‌های ترکیبی نظیر معماری‌های ترکیبی از شبکه‌های بازگشـتی و هم‌آمیختی به‌طـور موثـری تجزیه‌وتحلیل داده‌های مکانی و زمانی را ترکیب می‌کنند و آن‌ها را برای کارهایی ماننـد آنچه در کارهای [٤٤-٤٦] بررسی شـده اسـت، مناسب می‌سازند. با این حال، یک نقطه ضعف قابل توجه چنین مدل‌های ترکیبی افزایش پیچیدگی محاسباتی و زمان آموزش آن‌ها اسـت که ممکن است در محیط‌های با محدودیت منابع، چالش‌هایی ایجاد کند.

مجموعه دیگری از رویکردهایی که معمولاً برای ناوبری،

6-Convolutional Neural Networks (CNNs)
7-Recurrent Neural Networks (RNNs)

برنامه‌ریزی مسیر و اجتناب از موانع استفاده می‌شود، رویکرد یادگیری تقویتی عمیق است. یادگیری تقویتی عمیق به‌طور فزاینده‌ای برای دستیابی به تصمیم‌گیری مستقل در شرایط پیچیده استفاده می‌شود [۱۲, ۴۷, ۴۸] و عملکرد استثنایی را به‌ویژه در محیط‌های چالش‌برانگیز نشان داده است. رویکردهای یادگیری تقویتی عمیق قدرت بازنمایی یادگیری عمیق را با قابلیت یادگیری تقویتی برای یادگیری سیاست‌های بهینه از طریق آزمون‌وخطا ترکیب می‌کند. بدین‌صورت پهپادها را قادر می‌سازد تا با یادگیری رفتارهای پیچیده، نسبت به محیط عملیاتی خود سازگار شوند [۴۹].

یادگیری تقویتی به روشی از یادگیری اشاره دارد که در آن یک عامل درگیر فرآیند آزمون‌وخطا می‌شود و با دریافت سیگنال‌های پاداش از محیط بازخورد کسب می‌کند [۵۰]. هیچ مجموعه داده از پیش تعریف شده‌ای در این روش یادگیری وجود ندارد که مدل بخواهد از آن بیاموزد. در عوض، عامل داده‌های آموزشی را در زمان واقعی از طریق تعامل با محیط تولید می‌کند [۱۴]. یادگیری تقویتی عمیق این روش را با استفاده از شبکه‌های عصبی عمیق برای تخمین سیاست، توابع ارزش یا هر دو، برای تعیین سودمندترین انتخاب‌های رفتاری انجام می‌دهد.

در ناوبری پهپاد، یادگیری تقویتی عمیق به‌طور قابل‌توجهی در اجتناب از موانع، برنامه‌ریزی مسیر، انطباق بیدرنگ با شرایط پویا و هدایت دسته‌جمعی پهپادها موثر بوده است [۵۱، ۵۲]. به‌طور قابل‌توجهی، مطالعه [۵۳] بر توانایی یادگیری تقویتی عمیق برای مدیریت موثر مشاهدات حالت با ابعاد بالا تأکید می‌کند. علاوه بر این، یک مخزن حافظه ابتکاری را معرفی می‌کند که کارایی یادگیری را افزایش می‌دهد و آموزش را تسریع می‌بخشد و یادگیری تقویتی عمیق را به‌عنوان راه‌حلی قوی برای پهپادهایی که در محیط‌های پیچیده و غیرقابل پیش‌بینی کار می‌کنند، بر می‌شمارد.

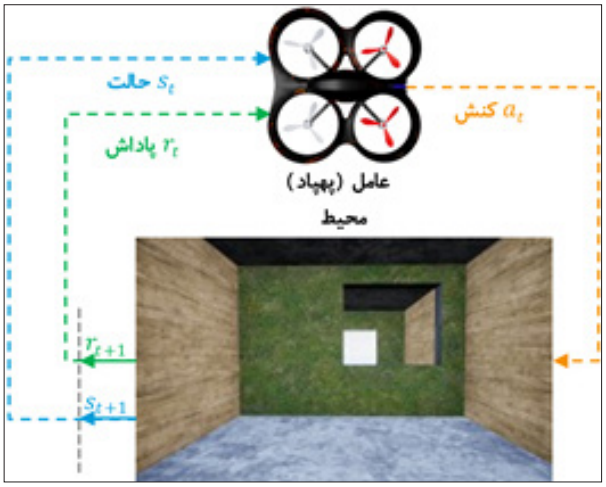
به‌عنوان مثال دیگری، [۵۴] یک رویکرد یادگیری تقویتی عمیق مبتنی بر حافظه را معرفی می‌کند که پهپادها را قادر می‌سازد از موانعی مانند عابران پیاده با دانش محیطی محدود اجتناب کنند. به‌طور مشابه، مطالعه [۵۵] بر روی ناوبری پهپاد با استفاده از دوربین تک‌چشمی و یادگیری عمیق تمرکز دارد و استفاده از حسگرهای مبتنی بر بینایی را برای ناوبری خودمختار و اجتناب از مانع برجسته می‌کند. به همین ترتیب، [۵۶] ناوبری پهپاد خودمختار را در محیط‌های پیچیده در مقیاس بزرگ با استفاده از یادگیری تقویتی عمیق مورد مطالعه قرار می‌دهد و کاربرد آن را برای ناوبری در شرایط چالش‌برانگیز در فضای باز نشان می‌دهد.

با وجود پیشرفت‌های قابل‌توجه، رویکردهای مورد بررسی محدودیت‌هایی دارند که استفاده از آن‌ها را در شرایط خاص مانند محیط‌های فاقد فناوری موقعیت‌یابی جهانی دشوار می‌سازد. بسیاری از این روش‌ها به چند حسگر یا دوربین‌های دوچشمی وابسته‌اند یا به دلیل طراحی‌های پیچیده، سربار محاسباتی زیادی دارند. وابستگی شدید به فناوری موقعیت‌یابی جهانی موجب کاهش کارایی این رویکردها در محیط‌هایی می‌شود که این فناوری موجود نیست یا سیگنال آن ضعیف است. در این مقاله، این محدودیت‌ها با استفاده از یک دوربین تک‌چشمی و حسگر واحد برای تصمیم‌گیری عملیاتی مبتنی بر داده‌های مشاهده شده برطرف می‌شود. این رویکرد نه تنها پیچیدگی‌های ناشی از تنظیمات سخت‌افزاری متعدد را از بین می‌برد بلکه به دلیل استفاده از حسگر ساده‌تر و پردازش داده‌های بصری، قابلیت تعمیم بیشتری در محیط‌های جدید دارد. مقایسه با روش‌های مرسوم و مدل‌های یادگیری عمیق پیچیده نشان می‌دهد که استفاده از چنین معماری سبکی، راه‌حلی عملی و مقیاس‌پذیر برای ناوبری خودمختار و اجتناب از موانع ارایه می‌دهد.

۳.روش پیشنهادی

در یادگیری تقویتی عامل در پاسخ به وضعیت فعلی

محیط تصمیم‌گیری می‌کند و برای انتخاب‌های خود بازخورد دریافت می‌کند. این فرآیند حلقه کنترل در شکل ۱ نشان داده شده است. مسایل یادگیری تقویتی معمولاً با استفاده از فرآیند تصمیم‌گیری مارکوف[8] تحت فرض محیط‌های کاملاً قابل مشاهده مدل می‌شوند. فرآیند تصمیم‌گیری مارکوف شامل چهار مؤلفه اصلی $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ است که به ترتیب نشان‌دهنده حالت‌ها، اقدامات، احتمالات انتقال و پاداش‌ها هستند.



شکل ۱. نمای سطح بالا از حلقه کنترل یادگیری تقویتی. عامل یک کنش را بر اساس حالت در محیط اعمال می‌کند، پاداش دریافت می‌کند و حالت بعدی را مشاهده می‌کند.

عامل به دنبال به حداکثر رساندن پاداش تجمعی در طول زمان است. پهپاد سیاست خود یعنی π را برای ارزش دادن به اقداماتی که پاداش کل مورد انتظار را بهینه می‌کند، اصلاح می‌کند. این هدف از طریق رابطه زیر [۱۴] بیان می‌شود:

$$\mathbb{E}_{s_t=s, a_t=a, \tau\sim\pi}[R_t] = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], \qquad (۱)$$

که در آن τ نشان‌دهنده یک مسیر نمونه‌برداری شده از سیاست π است، R_t پاداش تجمعی است که از زمان t شروع می‌شود، $\gamma \in [۰,۱)$ ضریب تنزیل است، و r_t نشان‌دهنده پاداش دریافت‌شده در مرحله زمانی t است.

8-Markov Decision Process (MDP)

چالش مهم آموزش عامل یادگیری تقویتی عمیق با الگوریتم‌های گرادیان سیاست، آسیب‌پذیری آن‌ها در برابر فروپاشی ناگهانی عملکرد است که موجب کاهش اثربخشی آن‌ها می‌شود [۵۷]. برطرف کردن این مشکل دشوار است چرا که عامل شروع به تولید مسیرهای غیر بهینه می‌کند که به‌عنوان داده‌های ضعیف برای به‌روزرسانی سیاست‌های بعدی عمل می‌کند و مشکل را پیچیده‌تر می‌کند.

بهینه‌سازی سیاست تقریبی[9] [۵۸] خانواده‌ای از رویکردهای بهینه‌سازی است که از یک معماری عملگر–نقاد برای رفع این چالش‌ها استفاده می‌کند. بهینه‌سازی سیاست تقریبی از یک رویکرد یادگیری مبتنی بر سیاست استفاده می‌کند که در آن سیاست با استفاده از مجموعه کوچکی از تجربیات جمع‌آوری شده از تعاملات با محیط به‌روز می‌شود. پس از به‌روزرسانی، این تجربیات کنار گذاشته شده و با استفاده از سیاست اصلاح‌شده، دسته جدیدی جمع‌آوری می‌شود.

بهینه‌سازی سیاست تقریبی یک معیار عملکرد سیاست نسبی را معرفی می‌کند که به‌طور کمّی میزان تفاوت عملکرد بین دو سیاست قدیم و جدید را اندازه‌گیری می‌کند. این معیار به شکلی طراحی شده است که تغییرات بزرگ در به‌روزرسانی سیاست را محدود کرده و به‌طور موثر تفاوت بین سیاست فعلی و سیاست جدید را کمیت‌سازی و کنترل می‌کند. با اعمال یک محدودیت در اندازه گام در طول به‌روزرسانی سیاست، از سقوط عملکرد جلوگیری می‌کند و بهبود یکنواخت را تضمین می‌کند. با $J(\pi)$ به‌عنوان تابع هدف، π نشان‌دهنده سیاست فعلی، و π' نشان‌دهنده سیاست به‌روز شده پس از یک تکرار، عملکرد نسبی سیاست را می‌توان به‌صورت زیر [۵۸] بیان کرد:

$$J(\pi') - J(\pi) = \mathbb{E}_{\tau\sim\pi'}\left[\sum_{t=0}^{T}\gamma^t A^{\pi}(s_t, a_t)\right], \qquad (۲)$$

که $A^{\pi}(s,a)$ تابع مزیت تحت سیاست π است که میزان بهتر یا بدتر بودن اقدامات عامل را با میانگین اقدام

9-Proximal Policy Optimization (PPO)

سیاست در یک وضعیت معین، ارزیابی می‌کند که به این صورت تعریف می‌شود [۵۷]:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s), \quad (۳)$$

که در آن $Q^{\pi}(s,a)$ تابع ارزش عمل را نشان می‌دهد و بازده مورد انتظار را توصیف می‌کند که از حالت s_t شروع می‌شود، اقدام a_t را انجام می‌دهد و سپس دنبال کردن سیاست π، تعریف شده به‌صورت (٤) است. $V^{\pi}(s)$ تابع ارزش حالت است که بازده مورد انتظار را با شروع از حالت $s_t s_t$ تحت سیاست π ارایه می‌کند که در (٥) بیان شده است [۵۷].

$$Q^{\pi}(s,a) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=.}^{\infty}\gamma^t r_t \mid s_. = s, a_. = a]. \quad (٤)$$

$$V^{\pi}(s) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=.}^{\infty}\gamma^t r_t \mid s_. = s]. \quad (٥)$$

یک رویکرد رایج در اجرای روش‌های گرادیان سیاست مبتنی بر معماری عملگر- نقاد شامل استفاده از برآورد مزیت تعمیم‌یافته[۱۰] [۵۹] است. این روش برای متعادل کردن تخمین‌های بازده کوتاه‌مدت و بلندمدت، مجموع نمایی خطاهای تفاضل زمانی[۱۱] را به‌صورت تنزیل و وزن شده معرفی می‌کند. این روش با میان‌یابی میان برآوردگرهای با سوگیری بالا و واریانس پایین (تفاضل زمانی تک‌مرحله‌ای) و برآوردگرهای با سوگیری کمتر و واریانس بیشتر (تفاضل زمانی n مرحله‌ای)، به توازن میان سوگیری و واریانس در برآورد مزیت می‌پردازد. این انعطاف‌پذیری اجازه می‌دهد تا مقادیر مزیت پایدارتر و دقیق‌تری محاسبه شود و به‌روزرسانی‌های سیاست بهبود پیدا کنند. از نظر ریاضی، برآورد مزیت تعمیم‌یافته به‌صورت زیر [۵۹] بیان می‌شود:

$$A^{\pi}_{\mathrm{GAE}}(s_t,a_t) = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}, \quad (٦)$$

که در آن $\lambda \in [۰,۱]$ مبادله بین سوگیری و واریانس را کنترل می‌کند. مقادیر کوچک‌تر λ تأکید بیشتری بر تخمین تابع ارزش دارند، در حالی که مقادیر بزرگ‌تر وزن بیشتری به پاداش‌های واقعی می‌دهند. δ_t در (٦) خطای تفاضل زمانی است [۱٤] که به‌صورت زیر تعریف می‌شود

$$\delta_t = r_t + \gamma V^{\pi}(s_{t+۱}) - V^{\pi}(s_t). \quad (۷)$$

تفاوت نسبی تعریف شده در (۲) به‌عنوان معیاری برای بهبود سیاست عمل می‌کند. تفاوت مثبت نشان می‌دهد که سیاست به‌روز شده، یعنی π'، بهتر از سیاست قبلی است. در هر تکرار سیاست، هدف انتخاب یک سیاست جدید π' است که این تفاوت را به حداکثر می‌رساند. در نتیجه، بهینه‌سازی هدف $J(\pi')$ معادل به حداکثر رساندن این شکاف عملکردی است، و هر دو را می‌توان از طریق صعود گرادیان[۱۲] به دست آورد؛ بنابراین می‌توانیم آن را به‌صورت زیر [۵۸] بیان کنیم:

$$\max_{\pi'} J(\pi') \Leftrightarrow \max_{\pi'}\big(J(\pi') - J(\pi)\big). \quad (۸)$$

بهینه‌سازی سیاست تقریبی در هسته خود از یک تابع هدف جایگزین استفاده می‌کند که بهبود سیاست ثابت را تضمین می‌کند که $J(\pi') - J(\pi) \geq ۰$ است. با این حال، محدودیت ایجاد می‌شود زیرا انتظارات (۲) نیازمند مسیرهایی از π' است که تا پس از به‌روزرسانی در دسترس نیست. برای پرداختن به این موضوع، فرض می‌کنیم که دو سیاست به هم نزدیک هستند، بنابراین توزیع حالت آن‌ها مشابه است. این کار اجازه می‌دهد تا رابطه را با استفاده از مسیرهای π با وزن‌های نمونه‌برداری با اهمیت۱۳ تقریب کنیم. می‌توانیم نسبت احتمال بین سیاست جدید و سیاست قدیمی را با پارامتر θ به‌صورت زیر [۵۸] بیان کنیم:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, \quad (۹)$$

که در آن $\pi_\theta(a_t|s_t)$ احتمال انجام عمل a_t در

10-Generalized Advantage Estimation (GAE)
11-Temporal Difference (TD)

12-Gradient Ascent
13-Importance Sampling

وضعیت s_t تحت سیاست جدید است و $\pi_{\theta_{old}}(a_t|s_t)$ احتمال تحت سیاست قدیمی است. بنابراین، می‌توانیم $r_t(\theta_{\text{old}}) = ۱$ را از آن دریافت کنیم. این اجازه می‌دهد تا تابع هدف (۲) را با مزایای محاسبه شده با استفاده از سیاست قدیمی به شرح زیر [۵۸] بازنویسی کنیم:

$$J(\theta) = \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_t^{\pi_{\theta_{old}}}(s_t, a_t)\right] = \mathbb{E}_t[r_t(\theta)A_t]. \quad (۱۰)$$

برای کنترل و جلوگیری از به‌روزرسانی‌های بزرگ و ناپایدار سیاست، یک نسخه بریده‌شده محدود از تابع جایگزین استفاده می‌شود که به‌روزرسانی‌های سیاست را محدود می‌کند تا در یک منطقه پایدار باقی بمانند که هم پیاده‌سازی آسان و هم از نظر محاسباتی کارآمد است [۵۷, ۵۸]:

$$J^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), ۱ - \epsilon, ۱ + \epsilon)A_t)], \quad (۱۱)$$

که در آن ϵ یک فراپارامتر است که همسایگی برش را تعریف می‌کند و می‌تواند در طول آموزش نزول یابد. این تابع برای اطمینان از به‌روزرسانی‌های محافظه‌کارانه، حداقل هدف‌های بریده نشده و بریده شده را می‌گیرد. بهینه‌سازی سیاست تقریبی سیاست را به‌طور مکرر با استفاده از صعود گرادیان [۱۴] به شرح زیر به‌روز می‌کند که α نشان‌دهنده نرخ یادگیری است:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J^{CLIP}(\theta). \quad (۱۲)$$

بهینه‌سازی سیاست تقریبی شامل یک هدف جداگانه برای آموزش تابع مقدار $V^\pi(s)$ با استفاده از خطای میانگین مربع است [۵۸]:

$$J_t^{VF}(\theta) = \mathbb{E}_t[(V_\theta(s_t) - R_t)^۲]. \quad (۱۳)$$

کاهش مقادیر ϵ در بریدن به تثبیت آموزش در الگوریتم با کاهش تدریجی انحراف مجاز کمک می‌کند. در ابتدا، مقادیر برش بزرگ‌تر آزادی بیشتری را برای کاوش و به‌روزرسانی‌های سریع‌تر می‌دهد، در حالی که کاهش ϵ به‌روزرسانی‌های محافظه‌کارانه‌تر را با پیشرفت آموزش تشویق می‌کند. برای محدوده برش ϵ، می‌توانیم از واپاشی خطی زیر استفاده کنیم:

$$\epsilon_t = \epsilon_{start} + (\epsilon_{end} - \epsilon_{start}) \times \left(۱ - \frac{t}{T}\right), \quad (۱۴)$$

که t مرحله زمانی فعلی از کل گام‌های زمانی T است، ϵ_t مقدار اپسیلون فعلی است، و ϵ_{start} و ϵ_{end} مقادیر شروع و پایان محدوده برش هستند.

کل تابع هزینه شامل تابع هدف سیاست، هزینه تابع ارزش و یک عبارت آنتروپی برای اطمینان از اکتشاف کافی است [۵۸]:

$$J_t^{TOTAL}(\theta) = \mathbb{E}_t[J_t^{CLIP}(\theta) - c_۱ J_t^{VF}(\theta) + c_۲ H(\pi_\theta)], \quad (۱۵)$$

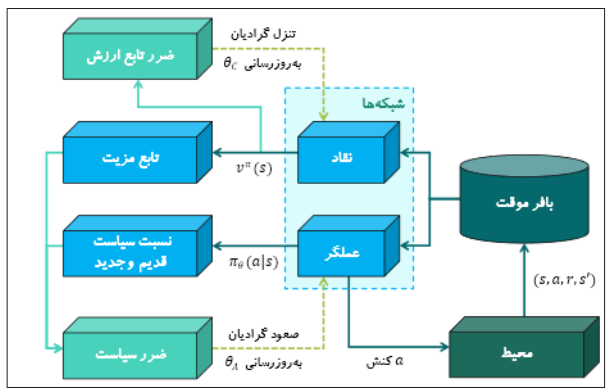
که در آن $c_۱$ و $c_۲$ ضرایبی برای کنترل اهمیت هر عبارت هستند و $H(\pi_\theta)$ آنتروپی سیاست برای اکتشاف است. نمودار جریان بهینه‌سازی سیاست تقریبی در شکل ۲ نشان داده شده است.

دوربین پهپاد تصاویر عمقی را با میدان دید ۹۰ درجه و ابعاد (۵۰، ۵۰، ۱) می‌گیرد. سپس داده را در محدوده [۱،۱] نرمال‌سازی می‌کنیم. ورودی از طریق لایه‌های پیچشی پردازش می‌شود، مسطح می‌شود و برای پردازش بیشتر به یک لایه کاملاً متصل وارد می‌شود. سپس ویژگی‌های خروجی به شبکه‌های عملگر و نقاد منتقل می‌شوند تا اقدامات و برآوردهای تابع ارزش را تعیین کنند. توجه به این نکته ضروری است که اقدامات مستقیماً خروجی نمی‌شوند. در عوض، شبکه عملگر میانگین و انحراف معیار یک توزیع نرمال[۱۴] را تولید می‌کند که از آن اقدامات نمونه‌برداری می‌شود. معماری روش پیشنهادی در شکل ۳ نشان داده شده است. با آموزش این شبکه، قابلیت درک

14-Normal Distribution

عمق و فاصله در تصاویر به دست خواهد آمد.

رویکرد پیشنهادی با استخراج اطلاعات فضایی و مکانی پیکسل‌ها، فرآیند تصمیم‌گیری مارکوف تک مرتبه‌ای را حفظ می‌کند. این رویکرد می‌تواند از طریق شدت نور هر پیکسل، میزان فاصله تا آن نقطه را درک کرده و از طریق روش تنزل گرادیان، خطاهای موجود را برطرف نماید. از آنجا که در هر گام زمانی اطلاعات فعلی برای تصمیم‌گیری کفایت می‌کند، نیاز به مدل‌سازی شبکه برای در نظر گرفتن وابستگی‌های کوتاه‌مدت نیست. با انجام این کار، ما نیاز به سازوکارهای بازگشتی غالباً با هزینه‌های گزاف محاسباتی، نظیر شبکه‌های عصبی بازگشتی را که وابستگی‌های بلندمدت را مدل‌سازی می‌کنند حذف می‌کنیم. با این رویکرد کارایی افزایش یافته و در عین حال یادگیری سیاست موثر آسان می‌شود. جزئیات پیاده‌سازی این روش در الگوریتم ۱ ارایه شده است.



شکل ۲. جریان داده الگوریتم بهینه‌سازی سیاست تقریبی.

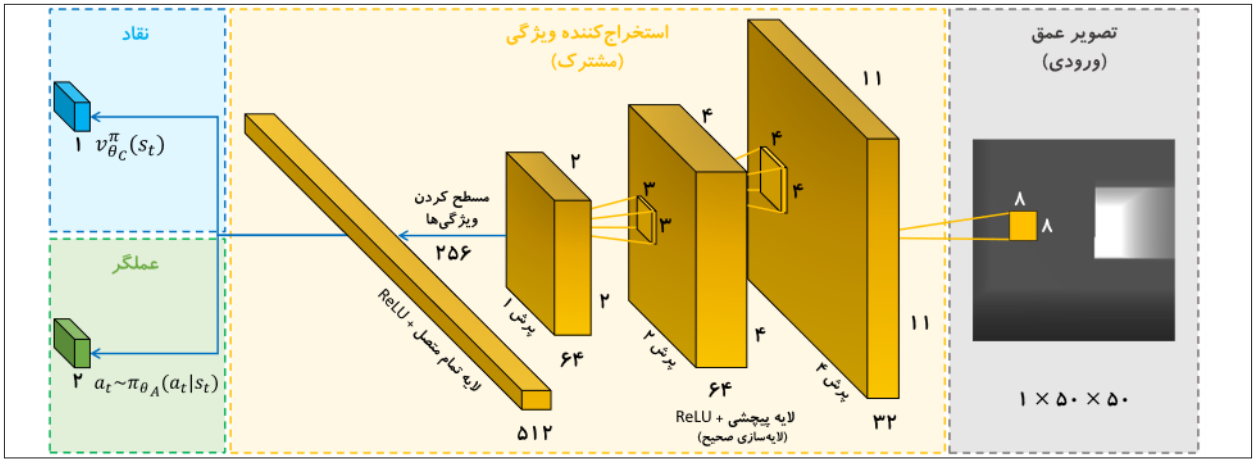
فضای عمل مسئله یک فضای کنترلی پیوسته دوبعدی و سطح بالا است که مستقیماً سرعت‌های خطی را در امتداد محورهای اصلی در سیستم مختصات محلی NED پهپاد مدیریت می‌کند. این سرعت‌ها برحسب متر بر ثانیه تعریف شده‌اند. حرکت در امتداد محور x روی یک مقدار ثابت تنظیم می‌شود و در طول پرواز بدون تغییر باقی می‌ماند. در نتیجه، پهپاد حرکت رو به جلو ثابت را حفظ می‌کند در حالی که مدل انتقال در امتداد محور y و انتقال ارتفاع در امتداد محور z را کنترل می‌کند. هر عمل به مدت نیم ثانیه روی پیشرانه اعمال می‌شود.

تابع پاداش پیشنهادی شامل سه عبارت است: اولین عبارت به عامل پس از گذراندن اولین گذرگاه بر اساس تعداد کل گذرگاه‌ها به‌طور مداوم به‌صورت زیر پاداش می‌دهد:

$$r_1 = \left\{ a + i \cdot \frac{b-a}{N-1} | i = 0,1,\dots,N-1 \right\}, \quad (16)$$

که در آن a مقدار شروع، b مقدار پایانی، N تعداد کل گذرگاه‌ها، و i شاخص گذرگاه فعلی است؛ عبارت دوم فاصله از مرکز گذرگاه فعلی را محاسبه می‌کند. هر گذرگاه یک مکعب فرضی با هر سه محور اصلی به طول دو متر است.

پهپاد بر اساس فاصله از مرکز مکعب پاداش منفی دریافت می‌کند و با نزدیک‌تر شدن به مرکز، جریمه آن



شکل ۳. معماری مدل پیشنهادی. تعداد فیلترها، پرش و اندازه خروجی برای هر لایه پیچشی ذکر شده است.

کمتر می‌شـــود. در مقابل، وقتی پهپاد داخل مکعب فرضی قرار می‌گیرد پاداش منـــفی به پاداش مثبت به‌صورت ذیل تبدیل می‌شود:

$$r_۲ = \begin{cases} \sqrt{\frac{۱}{\left|y_{\text{پهپاد}} - \frac{y_{\text{مکعب}}}{۲}\right|} + \frac{۱}{\left|z_{\text{پهپاد}} - \frac{z_{\text{مکعب}}}{۲}\right|}}, & \text{در داخل} \\ -\left(\sqrt{\|p_a - p_b\|}\right), & \text{در خارج} \end{cases} \quad (۱۷)$$

کـــه در آن $p_a = (y_{\text{پهپاد}}, z_{\text{پهپاد}})$ و $p_b = (y_{\text{مکعب}}, z_{\text{مکعب}})$ به ترتیب موقعیت پهپاد و موقعیت مکعب فرضی را نشـــان می‌دهند؛ سومین عبارت تابع پاداش به ازای گذر با موفقیت از تمام گذرگاه‌ها یک پاداش ثابـــت بزرگ‌تر ارایه می‌دهد؛ اما در صورت داشـــتن برخـــورد، جریمه‌ای در نظر گرفته می‌شـــود و دوره خاتمـــه می‌یابد. پاداش نهـــایی به‌عنوان مجموع تمام عبارات به‌صورت ذیل محاسبه می‌شود:

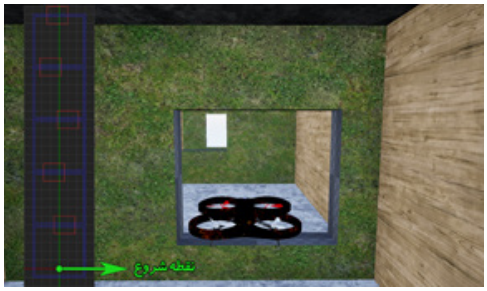
$$r_t = r_۱ + r_۲ + r_۳. \quad (۱۸)$$

۴. نتایج تجربی

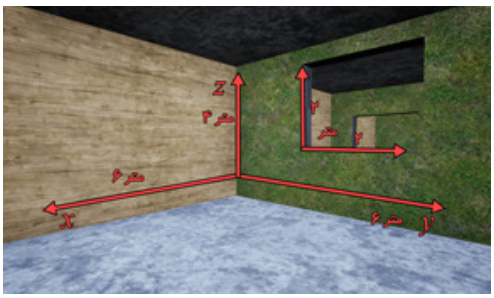
تمـــام مراحـــل آزمایـــشی روی یک رایانـــهٔ قابل حمل معمولی با مشـــخصات پردازنده Intel Core i7-6700HQ با فرکانـــس ۲/٦ گیگاهرتز، ١٦ گیگابایـــت حافظه و کارت گرافیک GeForce GTX 960M با ٤ گیگابایت حافظه داخلی انجام شده است. زبان برنامه‌نویسی پایتون ۳ بوده و برای پیاده‌سازی شبکه‌های عصبی عمیق از کتابخانه پایتورچ در سیستم عامل ویندوز ۱۰ استفاده شده است.

برای شبیه‌ســـازی واقع‌گرایانه، از ســـکوی شبیه‌ساز ایرسیم[15] [٦٠] استفاده شده که به‌منظور توسعه الگوریتم‌های یادگیری تقویتی با قابلیت اجرا در دنیای واقعی طراحی شده است. این شبیه‌ساز با استفاده از موتور گرافیکی غیرواقعی، ویژگی‌های محیطی و آیرودینامیکی پهپاد را ایجاد می‌کند. در این پژوهش، تمرکز بر استفاده از پهپاد چندموتوره است که به دلیل قابلیت‌های شناوری، مانورپذیری و کاربردهای غیرنظامی گسترده‌اش شناخته شده است.

15-AirSim



شکل ۴. محیط آموزش و نقشه آن

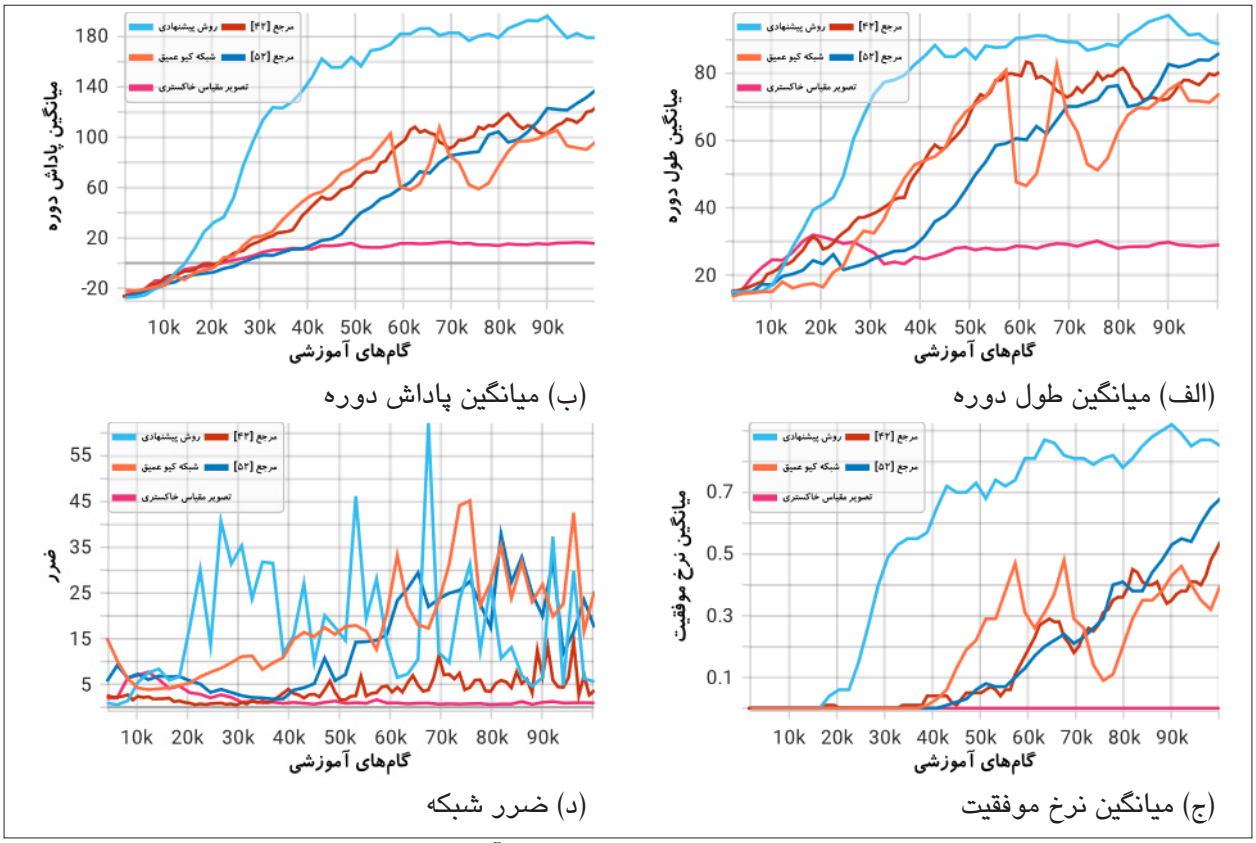


شکل ۵. ابعاد هر بخش و گذرگاه در محیط.

محیط بر اســـاس کار [٦١] طـــراحی و به چند بخش تقسیم شـــده اســـت. یک دیوار بـــا گـــذرگاه باریک هر بخـــش را جدا می‌کند و پهپاد باید بـــرای ادامه مسیر از آن عبور کند. بخـــش نهایی، پهپاد را از محیط داخلی به محیط بیرونی منتقل می‌کند کـــه این انتقال پایان وظیفه را مشـــخص می‌نماید. محیط آموزشی و نقشه مربوط به آن در شـــکل ٤ نشان داده شده است. همچنین، ابعاد عناصر موجود در محیط در شـــکل ٥ نشـــان داده شده است.

بـــرای مقایســـه، روش پیشـــنهادی را در برابـــر رویکردهای مرتبط نزدیک از قبیل [٤٢] و [٥٢] ارزیابی می‌کنیم. علاوه بر این، رویکرد یادگیری کیو عمیق[16] که در ســـال ۲۰۱۵ معرفی شـــده را به‌عنوان یکی دیگر از الگوریتم‌های اساسی در کنار الگوریتم پایه بهینه‌سازی سیاســـت تقریبی که تنها ورودی آن تک تصویر مقیاس خاکســـتری اســـت بررسی می‌کنیم. معیار اصلی برای ارزیابی یـــک الگوریتم ناوبری، میزان نـــرخ موفقیت و برخورد اســـت. در کنار آن‌ها، انـــدازه طول هر دوره و پـــاداش دریافتی در هر دوره نیـــز عوامل دیگری برای

16-Deep Q-Network



شکل ۶. نتایج عملکرد عامل در زمان آموزش

سنجش موفقیت یک مدل یادگیری تقویتی است. فراپارامترهای آموزشی در جدول ۱ به تفصیل آمده است. فراپارامترهای مدل از طریق مجموعه‌ای از آزمایش‌های مکرر و تحلیل دقیق عملکرد مدل در محیط تعیین شده‌اند. این آزمایش‌ها با هدف دستیابی به بهینه‌ترین عملکرد ممکن به خصوص در میزان موفقیت پهپاد انجام شده‌اند. دلیل انتخاب این مقادیر بر اساس نتایج حاصل از این شبیه‌سازی‌ها بوده و فرآیند تنظیمات با هدف ارایه مدل بهینه در این مقاله انجام شده است.

شکل ٦ نتایج عملکرد مدل در زمان آموزش را نشان می‌دهد. همان‌طور که از نمودارهای این شکل مشهود است، نمودارهای (الف)، (ب)، (ج) شباهت‌های قابل‌توجهی دارند و نمودارهای (الف) و (ب) همبستگی قوی را نشان می‌دهند. همان‌طور که آموزش در محیط پیشرفت می‌کند، اولین روند قابل مشاهده در این نمودارها این است که عامل به‌طور موثر از تعاملات در طول زمان یاد می‌گیرد و در نتیجه میانگین طول دوره در اکثر رویکردها افزایش می‌یابد. این نشان می‌دهد که عامل از طریق به‌روزرسانی‌های وزن یک سیاست موثر را می‌آموزد.

علاوه بر این، افزایش میانگین طول دوره با افزایش پاداش همراه است که یادگیری موثر را بیشتر برجسته می‌کند. در میان رویکردهای مقایسه شده در این شکل، مشهود است که هنگام استفاده از تصاویر منفرد در مقیاس خاکستری، بهبود قابل‌توجهی در میانگین طول دوره یا پاداش در مقایسه با روش‌های دیگر مشاهده نمی‌شود. این نشان می‌دهد که سیاست نیاز به درک معناداری برای خروجی دادن کنش مناسب است؛ بنابراین، از تک تصویر واحد و ویژگی‌های استخراج شده آن، سیاست درک درستی از حالت پیدا نکرده و منجر به شکست این روش می‌شود.

الگوریتم ۱. بهینه‌سازی سیاست تقریبی

ورودی: تصویر عمقی تک کانال با اندازه (C, H, W)

خروجی: اقدام پیوسته $a_t a_t$

تنظیم $c_2 \geq 0$، وزن همگن‌سازی آنتروپی

تنظیم $c_1 \geq 0$، وزن اهمیت $J_t^{VF}(\theta)$

تنظیم $\epsilon \geq 0$، متغیر برش

تنظیم K، تعداد دوره‌ها

تنظیم T، افق زمان

تنظیم $M \leq T$، دسته کوچک

تنظیم $\alpha \geq 0$، نرخ یادگیری

مقداردهی اولیه متعامد پارامترهای وزن و سوگیری عملگر-نقاد θ_A, θ_C

مقداردهی شبکه عملگر $\theta_{A_{old}}$

به ازای کل قدمها، $i = 1, 2, \ldots$ تکرار کن

  ایجاد یک بافر خالی

  دریافت حالت کنونی $s_t s_t$

  به ازای $j = 1, 2, \ldots, T$ تکرار کن

    انتخاب و اعمال کنش $a_t \sim \pi_{\theta_A}(a_t|s_t)$

    دریافت r_t و حالت بعدی s_{t+1}

    ثبت (s_t, a_t, r_t, s_{t+1}) در بافر موقت

    $s_{t+1} \rightarrow s_t$

  پایان حلقه

  تنظیم $\theta_{A_{old}} = \theta_A$

  محاسبه $V_{tar,1}^{\pi}, \ldots, V_{tar,T}^{\pi}$ با شبکه نقاد و پارامتر θ_C

  محاسبت مزیت $A_1, \ldots, A_T$ با استفاده از (۶) با $\theta_{A_{old}}$

  به ازای K دوره $= 1, 2, \ldots,$ تکرار کن

    به ازای دسته کوچک m در کل دسته تکرار کن

      محاسبه $r_m(\theta_A)$

      محاسبه $J_m^{CLIP}(\theta_A)$ با (۱۱)، با مزیت A_m و $r_m(\theta_A)$

      محاسبه آنتروپی H_m با $\theta_A \theta_A$

      محاسبه ضرر سیاست $J_{pol}(\theta_A) = J_m^{CLIP}(\theta_A) - c_2 H_m$ طبق (۱۵)

      محاسبه ارزش‌های تخمینی $v^{\pi}(s_m)$ با θ_C

      محاسبه ضرر تابع ارزش $J_{val}(\theta_C)$ با استفاده از (۱۳)

      به‌روزرسانی پارامترهای عملگر θ_A با(۱۲) و $J_{pol}(\theta_A)$

      به‌روزرسانی پارامترهای منتقد θ_C با(۱۲) و $J_{val}(\theta_C)$

    پایان حلقه

  پایان حلقه

  کاهش بازه برش متغیر $\epsilon\epsilon$

پایان حلقه

با توجه به این که کنش‌ها پیوسته هستند، مشخص می‌شود که در نظر گرفتن همبستگی بین تصاویر متعدد برای ثبت تاثیر تغییرات کنش بر فضای حالت ورودی ضروری است؛ اما این در صورتی است که تصویر ورودی اطلاعات معناداری را به‌صورت ذاتی منتقل نکند. در رویکرد پیشنهادی به حل این مسئله پرداخته شده است و از تصاویر عمقی به‌عنوان ورودی فضای حالت استفاده شده است. تصاویر عمقی اطلاعات فاصله را در هر پیکسل رمزگذاری می‌کند. از طریق استخراج ویژگی‌های این تصویر به‌واسطه لایه‌های پیچشی، شبکه توانایی درک معنادار شدت نور هر پیکسل را پیدا می‌کند. همان‌طور که از نمودارها به خصوص نمودار (ج) مشهود است، رویکرد پیشنهادی از نظر میانگین طول دوره ۲/۷ گام یا ۳/۱ درصد بهبود، ٤١/٢٢ از لحاظ میانگین پاداش دریافتی یا معادل ۲۹/۹ درصد بهبود، و ۰/۱۷ معادل با ۲۵/۰ درصد نرخ موفقیت بیشتر نسبت به بهترین رویکرد از ما بین رویکردهای مرجع قیاس شده، بهترین نتایج را دریافت کرده است.

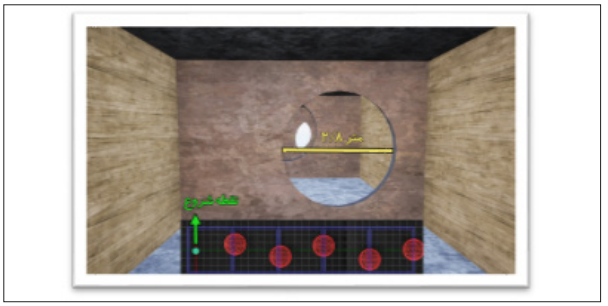
با این حال، در نمودار (د) به‌طور قابل‌توجهی دیده می‌شود که رویکرد پیشنهادی دارای نوسانات متعددی است. علت این امر آن است که شبکه در هرلحظه، به دلیل دریافت ورودی‌های جدید که با ورودی‌های مشاهده‌شده قبلی متفاوت هستند، سعی در بهبود تابع ضرر دارد. به دلیل افزایش پاداش دریافتی، مشاهده می‌شود که این نمودار نوسانات بیشتری را تجربه می‌کند.

در یادگیری تقویتی، عموماً زمانی که همگرایی زودهنگام رخ می‌دهد، خطای شبکه، بخصوص در مراحل ابتدایی، به‌سرعت کاهش می‌یابد. در مقابل، زمانی که شبکه با اطلاعات جدیدی مواجه می‌شود که پیش‌تر آن‌ها را تجربه نکرده است، ضرر بیشتری تحمیل می‌شود، چرا که شبکه این اطلاعات را پیش‌بینی نکرده است. این موضوع به‌وضوح در رویکرد تک

تصویـــر با مقیـــاس خاکســـتری کاملاً از نمودارها نیز دریافت می‌شود.

جدول ۱. فراپارامترهای آموزشی.

فراپارامتر	مقدار
بهینه‌ساز	آدام
دانه	۲۰۲۴
نرخ یادگیری	۰۰۰۰۷/۰
گام‌های جمع‌آوری داده	۲۰۴۸
اندازه دسته	۱۲۸
تعداد تکرار هر به‌روزرسانی	۳۰
ضریب تنزیل	۰/۹۹
ضریب برآورد مزیت تعمیم‌یافته	۰/۷
حداکثر نرم گرادیان	۰/۵
برش شروع	۰/۳
برش نهایی	۰/۰۵
تعداد گام‌های یادگیری	۱۰۰/۰۰۰



شکل ۷. پیکربندی و نقشه آزمایش دوم تعمیم‌پذیری

با وجود نوســـانات نمودار ضرر، رویکرد پیشـــنهادی توانسته است با به‌روزرسانی کنترل‌شده از طریق برش و محدودسازی نرم گرادیان، در سه معیار دیگر که مهم‌ترین آن‌ها موفقیت عامل است، ثبات و پایداری را حفظ کند تا از تغییرات سریع در سیاست اجتناب کرده و بهترین عملکرد را در تمام معیارها به دست آورد. رویکردهای [۴۲] و [۵۲] عملکرد بسیار ضعیف‌تری در حل مســـئله و دســـتیابی به نرخ‌های موفقیت نشـــان می‌دهند و ایـــن موضوع اعتبار و اثربخشی روش پیشنهادی را بیشتر تایید می‌کند. الگوریتم یادگیری کیو عمیق نیز عملکرد به‌مراتب ضعیف‌تری در حل این مســـئله از خود نشان می‌دهد و عموماً دچار فروپاشی سیاســـت و یادگیری می‌شـــود؛ چرا که برخلاف الگوریتم پیشنهادی، ســـازوکاری برای برش تغییرات شدید ندارد، هرچند این رویکردِ به‌روزرسانی معیوب را تا حدودی در طول آموزش جبران می‌کند.

جزئیات کامل آزمایش‌های انجام‌شده در جدول ۲ نشان داده شده است. بر اساس این جدول، در معیار میانگین طول دوره رویکرد پیشـــنهادی ۲۰۶/۳ درصد نسبت به رویکرد تک تصویر، ۱۹/۹ درصد نســـبت به رویکرد شـــبکه کیو، ۳/۱ درصد نســـبت به رویکرد مرجع [۴۲] و ۱۰/۴ درصد نسبت به رویکرد مرجع [۵۲] بهبود حاصل شده است. در معیار میانگین پاداش دوره ۱۰۵۳/۲ درصد بهبود نسبت به رویکرد تک تصویر، ۸۵/۰ درصد نســـبت به یادگیری کیو عمیق، ۲۹/۹ درصد نســـبت به مرجع [۴۲] و ۴۴/۵ درصد نســـبت به مرجع [۵۲] بهبود داشـــته اســـت. در معیار نرخ موفقیت نیز ۱۱۲/۵ درصد نسبت به رویکرد یادگیری کیو، ۲۵/۰ درصد نســـبت به مرجع [۴۲] و ۵۷/۴ درصد نسبت به مرجع [۵۲] بهبود وجود داشته است. در سنجش میزان ضرر شـــبکه نســـبت به بهترین رویکرد از لحـــاظ میزان موفقیـــت که مرجع [۴۲] اســـت، رویکرد پیشـــنهادی ۶۷/۸ درصد بهبود داشته و توانســـته است میزان خطای شبکه را بهتر کاهش دهد.

بـــرای ارزیابی توانـــایی تعمیم مـــدل دو آزمون انجام می‌دهیم. در آزمون اول موقعیت گذرگاه‌ها را تغییر می‌دهیم و در آزمون دوم هم موقعیت و هم‌شـــکل آن‌ها را از مکعب به کره تغییر می‌دهیم. شـــکل ۷ آزمایش دوم تعمیم‌پذیری و نقشه آن را نشان می‌دهد. رویکرد پیشنهادی در هر سه مورد فوق‌العاده عمل می‌کند که نشان‌دهنده توانایی تعمیم قوی مدل اســـت. مدل با موفقیت می‌آمـــوزد که چگونه از طریق اســـتنتاج از حالت‌های ورودی، بـــه مرکز گذرگاه‌ها حرکت کند. فرآیند یادگیـــری موفق، اثربخشی تابع پاداش را در مرحلـــه آموزش نمایش می‌دهد که موجب اســـتنتاج عالی مســـئله در مرحله آزمایش می‌شـــود. این مورد حتی زمانی که نقشـــه محیط تغییر یافته و یا شکل گذرگاه تغییر می‌یابد نیز برقرار اســـت. نتایج حاصله نشـــان می‌دهد که رویکرد پیشـــنهادی می‌تواند در پهپادها استفاده شود و به

جدول ۲. نتایج عملکرد مدل در مرحله آموزش.

رویکرد	ضرر شبکه	نرخ برخورد (٪) ↓	نرخ موفقیت (٪) ↑	میانگین پاداش دوره ↑	میانگین طول دوره ↑
تک تصویر	۰/۹۷	۱۰۰	۰	۱۵/۴۳	۲۸/۹۵
شبکه کیو	۲۵/۳۷	۰/۶۰	۰/۴۰	۹۶/۹۱	۷۳/۹۱
مرجع [۴۲]	۱۷/۴۴	۰/۳۲	۰/۶۸	۱۳۷/۹۶	۹۸/۸۵
مرجع [۵۲]	۳/۶۹	۰/۴۶	۰/۵۴	۱۲۴/۰۰	۲۹/۸۰
پیشنهادی	۵/۶۰	۰/۱۵	۰/۸۵	۱۷۹/۱۸	۸۸/۶۸

جدول ۳. نتایج عملکرد عامل در مرحله آزمایش و ارزیابی تعمیم‌پذیری مدل طی ۱۰۰ تکرار

دسته‌بندی آزمایش	میانگین طول دوره	میانگین پاداش دوره	نرخ موفقیت (٪)	نرخ برخورد (٪)
اصلی	۹۸	۲۴۶/۱۷	۱۰۰	۰
تعمیم‌پذیری اول	۹۸	۲۳۰/۶۷	۱۰۰	۰
تعمیم‌پذیری دوم	۹۷/۵۲	۲۱۴/۳۲	۱۰۰	۰

آن‌ها توانـــایی ناوبری و مسیریابی از طریق درک تصویر و در نهایـــت انجام کنش به‌وسیلـــه یادگیری تقویتی عمیق را ارایـــه کند. نتایج جامع همه آزمایش‌ها در جدول ۳ ارایه شده است.

۵. جمع‌بندی

در این مقاله یک رویکرد نوآورانه برای ناوبری خودکار پهپادهـــا در گذرگاه‌هـــای باریک با اســـتفاده از یادگیری تقویتی عمیق و پردازش تصاویـــر عمقی ارایه گردید. این سیستم با اجتناب از وابستگی به حسگرهای گران‌قیمت و حجیم، امکان ناوبری موثر تنها با داده‌های تصویری عمق را نشـــان می‌دهد. نتایج آزمایش‌ها توانایی سیســـتم را در کاهش برخوردها و دســـتیابی به‌دقت بالا در ناوبری تایید کرده و بهبود قابل‌توجهی نسبت به روش‌های سنتی نشان می‌دهد.

یـــکی از ویژگی‌هـــای برجســـته رویکرد پیشـــنهادی، تعمیم‌پذیـــری بـــالای آن در تنظیمـــات و پیکربندی‌هـــای گوناگـــون محیطی اســـت. مدل ارایه شـــده با اســـتفاده از شبکه‌های عصبی پیچشی برای استخراج ویژگی‌های عمق و یادگیری تقویتی برای تصمیم‌گیری، قادر به سازگاری با تغییرات محیطی نظیر شـــکل گذرگاه‌ها و تنوع مکان آن‌ها است. این تعمیم‌پذیری نشان‌دهندهٔ پتانسیل این روش برای استفاده در محیط‌های ناشناخته و غیرقابل پیش‌بینی است؛ بخصوص برای مکان‌هایی که انعطاف‌پذیری بالا از پهپادها انتظار می‌رود. علاوه بر این، روش پیشـــنهادی با معماری ساده و ســـبک‌تر، بار محاســـباتی کمتری نسبت به سایر سامانه‌های ناوبری دارد و همین امر موجب افزایش کاربرد عملی آن می‌شـــود. این پژوهش مبنای توسعهٔ راه‌حل‌های مقرون‌به‌صرفـــه و مقیاس‌پذیر برای ناوبـــری پهپادها در محیط‌هایی با منابع محدود را فراهم می‌کند.

برای جهت‌های آینده، توســـعه الگوریتم‌های اجتناب از موانع، مانند برنامه‌ریزی مسیر به سمت یک هدف مشخص در شـــرایط محیطی پویا و پیچیده‌تر، نظیر موانع متحرک، مدنظر اســـت تا کاربرد رویکرد فعلی برای مسایل نیازمند به بیدرنگ گســـترش یابد. علاوه بر این، مطالعات تطبیقی بـــا الگوریتم‌های پیشـــرفته در وضعیت‌های مختلف انجام خواهد شد.

۶. مراجع

[1] N. Elmeseiry, N. Alshaer, T. Ismail, "A Detailed Survey and Future Directions of Unmanned Aerial Vehicles (UAVs) with Potential Applications," Aerospace, vol. 8, no. 12, 2021.
[2] M. A. Tahir, I. Mir, T. U. Islam, "A Review of UAV Platforms for Autonomous Applications: Comprehensive Analysis and Future Directions," IEEE Access, vol. 11, pp. 52540-

52554, 2023.
[3] S. A. H. Mohsan, M. A. Khan, F. Noor, I. Ullah, M. H. Alsharif, “Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive Review,” Drones, vol. 6, no. 6, 2022.
[4] D. J. Yeong, G. Velasco-Hernandez, J. Barry, J. Walsh, “Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review,” Sensors, vol. 21, no. 6, 2021.
[5] K. Telli et al., “A Comprehensive Review of Recent Research Trends on Unmanned Aerial Vehicles (UAVs),” Systems, vol. 11, no. 8, 2023.
[6] S. Campbell et al., “Sensor Technology in Autonomous Vehicles : A review,” 2018, pp. 1-4.
[7] M. Kim, J. Kim, M. Jung, H. Oh, “Towards monocular vision-based autonomous flight through deep reinforcement learning,” Expert Systems with Applications, vol. 198, p. 116742, 2022.
[8] J. Li, X. Xiong, Y. Yan, Y. Yang, “A Survey of Indoor UAV Obstacle Avoidance Research,” IEEE Access, vol. 11, pp. 51861-51891, 2023.
[9] S. Rezwan, W. Choi, «Artificial Intelligence Approaches for UAV Navigation: Recent Advances and Future Challenges,” IEEE Access, vol. 10, pp. 26320-26339, 2022.
[10] B. Mahdipour, S. H. Zahiri, I. Behravan, “An Intelligent Two and Three Dimensional Path Planning, Based on a Metaheuristic Method,” Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 13, no. 1, pp. 93-116, 2025.
[11] S. Y. Choi, D. Cha, “Unmanned aerial vehicles using machine learning for autonomous flight; state-of-the-art,” Advanced Robotics, vol. 33, no. 6, pp. 265-277, 2019.
[12] A. Carrio, C. Sampedro, A. Rodriguez-Ramos, P. Campoy, “A Review of Deep Learning Methods and Applications for Unmanned Aerial Vehicles,” Journal of Sensors, vol. 2017, no. 1, p. 3296874, 2017.
[13] Y. Chang, Y. Cheng, U. Manzoor, J. Murray, “A review of UAV autonomous navigation in GPS-denied environments,” Robotics and Autonomous Systems, vol. 170, p. 104533, 2023.
[14] R. S. Sutton, “Reinforcement learning: An introduction,” A Bradford Book, 2018.
[15] H. Taheri, S. Rasoul Hosseini, M. A. Nekoui, “Deep Reinforcement Learning with Enhanced PPO for Safe Mobile Robot Navigation,” arXiv e-prints, p. arXiv:2405.16266, 2024.
[16] A. S. Sadr, M. S. Khojasteh, H. Malek, A. Salimi-Badr, “An Efficient Planning Method for Autonomous Navigation of a Wheeled-Robot based on Deep Reinforcement Learning,” in 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), 2022, pp. 136-141.
[17] S. Mashhouri, M. Rahmati, Y. Borhani, E. Najafi, “Reinforcement Learning based Sequential Controller for Mobile Robots with Obstacle Avoidance,” in 2022 8th International Conference on Control, Instrumentation and Automation (ICCIA), 2022, pp. 1-5.
[18] S. Sabzekar, M. Samadzad, A. Mehditabrizi, A. N. Tak, “A Deep Reinforcement Learning Approach for UAV Path Planning Incorporating Vehicle Dynamics with Acceleration Control,” Unmanned Systems, vol. 12, no. 03, pp. 477-498, 2024.
[19] H. S. M. Mahalegi, A. Farhadi, G. Molnár, E. Nagy, “Enhancing UAV Autonomous Navigation in Indoor Environments Using Reinforcement Learning and Convolutional Neural Networks,” in 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), 2024, pp. 000091-000100.
[20] M. Ramezani, M. A. Amiri Atashgah, A. Rezaee, “A Fault-Tolerant Multi-Agent Reinforcement Learning Framework for Unmanned Aerial Vehicles–Unmanned Ground Vehicle Coverage Path Planning,” Drones, vol. 8, no. 10, p. 537, 2024.
[21] P. R. Gervi, A. Harati, S. K. Ghiasi-Shirazi, “Vision-Based Obstacle Avoidance in Drone Navigation using Deep Reinforcement Learning,” in 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), 2021, pp. 363-368.
[۲۲] م. دین پرســت، ع. رودباری، «کنترل لندینگ پهپاد با دینامیک نامشخص بــا اســتفاده از یادگیری تقویــتی،» بیســت و یکمین کنفرانس مــلی مهندسی برق،کامپیوتر و مکانیک، ۱۴۰۳.
[۲۳] ج. روشــنی یان، ف. خواجه محمدی، «تبیین و پیاده سازی ۴ روش یادگیری تقویتی هوش مصنوعی (برنامه ریزی پویا، مونت کارلو، تفاضلات زمانی (سارســا و یادگیری)) برای مسیریابی یک کوادروتور در حضور موانع در صفحه با فرض گسســته سازی،» بیست و دومین کنفرانس بین المللی انجمن هوافضای ایران، ۱۴۰۲.
[۲۴]م. خاکبــاز، م. انجیــدنی، «یادگیــری مسیر مناســب و هدایــت اتوماتیک کوادروتور،» پنجمین همایش مــلی فناوریهای نوین در مهندسی برق، کامپیوتر و مکانیک ایران، ۱۴۰۱.
[۲۵]ا. شــریفی، آ. الســتی، «طراحی کنترل PID خودتنظیم با یک ساختار عصبی مبتنی بر عملگر-منتقد بــرای کنترل وضعیت و ارتفــاع کوادروتور،» سی امین همایش سالانه بین المللی انجمن مهندسان مکانیک ایران، ۱۴۰۱.
[26] H. Alvarez, L. M. Paz, J. Sturm, D. Cremers, «Collision Avoidance for Quadrotors with a Monocular Camera,” in Experimental Robotics: The 14th International Symposium on Experimental Robotics, M. A. Hsieh, O. Khatib, and V. Kumar Eds. Cham: Springer International Publishing, 2016, pp. 195-209.
[27] G. Cho, J. Kim, H. Oh, “Vision-Based Obstacle Avoidance Strategies for MAVs Using Optical Flows in 3-D Textured Environments,” Sensors, vol. 19, no. 11, p. 2523, 2019.
[28] W. E. Green, P. Y. Oh, “Optic-Flow-Based Collision Avoidance,” IEEE Robotics & Automation Magazine, vol. 15, no. 1, pp. 96-103, 2008.
[29] A. Eresen, N. İmamoğlu, M. Önder Efe, “Autonomous quadrotor flight with vision-based obstacle avoidance in virtual environment,” Expert Systems with Applications, vol. 39, no. 1, pp. 894-905, 2012.
[30] S. Aggarwal, N. Kumar, “Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges,” Computer Communications, vol. 149, pp. 270-299, 2020.
[۳۱] س. م. حسینی رســتمی، ح. خالوزاده، م. کمــارجی، «اجتناب از موانع ربات سیار با اســتفاده از الگوریتم میدان پتانسیل مجازی اصلاح شده،» علوم رایانشی، جلد ۲، شماره ۱، صفحات ۴۵-۳۴، ۲۰۱۷.
[32] B. Y. Li, H. Lin, H. Samani, L. Sadler, T. Gregory, B. Jalaian, “On 3D autonomous delivery systems: Design and development,” 2017, pp. 1-6.

[33] A. Puente-Castro, D. Rivero, A. Pazos, E. Fernandez-Blanco, "A review of artificial intelligence applied to path planning in UAV swarms," Neural Computing and Applications, vol. 34, no. 1, pp. 153-170, 2022.
[34] J. Gao, Y. Zheng, K. Ni, Q. Mei, B. Hao, L. Zheng, «Fast Path Planning for Firefighting UAV Based on A-Star algorithm," Journal of Physics: Conference Series, vol. 2029, no. 1, p. 012103, 2021.
[35] L. Liu, X. Wang, X. Yang, H. Liu, J. Li, P. Wang, "Path planning techniques for mobile robots: Review and prospect," Expert Systems with Applications, vol. 227, p. 120254, 2023.
[36] K. N. McGuire, G. C. H. E. de Croon, K. Tuyls, "A comparative study of bug algorithms for robot navigation," Robotics and Autonomous Systems, vol. 121, p. 103261, 2019.
[37] N. Mahdian, S. H. Attarzadeh-Niaki, A. Salimi-Badr, "A Systematic Embedded Software Design Flow for Robotic Applications," in 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), 2021, pp. 217-222.
[38] M. Vazirpanah, S. H. Attarzadeh-Niaki, A. Salimi-Badr, "ROS-Based Co-Simulation for Formal Cyber-Physical Robotic System Design," in 2022 27th International Computer Conference, Computer Society of Iran (CSICC), 2022, pp. 1-5.
[39] Y. Guo, X. Liu, X. Liu, Y. Yang, W. Zhang, "FC-RRT*: An Improved Path Planning Algorithm for UAV in 3D Complex Environment," ISPRS International Journal of Geo-Information, vol. 11, no. 2, 2022.
[40] T. Elmokadem, A. V. Savkin, "Towards Fully Autonomous UAVs: A Survey," Sensors, vol. 21, no. 18, p. 6223. 2021.
[۴۱] ف. مولایی، ع. موســوی، م. دولتشاهی، «طراحی پهپاد مسیریاب هوشمند با استفاده از یادگیری عمیق و منطق فازی،» سیستم های فازی و کاربردها، جلد ۷، شماره ۱، صفحات ۲۰۷-۱۸۹، ۲۰۲۴.
[42] S. Y. Shin, Y. W. Kang, Y. G. Kim, "Automatic Drone Navigation in Realistic 3D Landscapes using Deep Reinforcement Learning," in 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), 2019, pp. 1072-1077.
[43] S. Chehelgami, E. Ashtari, M. A. Basiri, M. Tale Masouleh, A. Kalhor, "Safe deep learning-based global path planning using a fast collision-free path generator," Robotics and Autonomous Systems, vol. 163, p. 104384, 2023.
[44] R. J. Alitappeh, N. Mahmoudi, M. R. Jafari, A. Foladi, "Autonomous Robot Navigation: Deep Learning Approaches for Line Following and Obstacle Avoidance," in 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), 2024, pp. 1-6.
[45] L. O. Rojas-Perez, J. Martinez-Carranza, "DeepPilot: A CNN for Autonomous Drone Racing," Sensors, vol. 20, no. 16, p. 4524, 2020.
[46] S. Daftry, S. Zeng, J. A. Bagnell, M. Hebert, "Introspective perception: Learning to predict failures in vision systems," in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 1743-1750.
[47] A. V. R. Katkuri, H. Madan, N. Khatri, A. S. H. Abdul-Qawy, K. S. Patnaik, "Autonomous UAV navigation using deep learning-based computer vision frameworks: A systematic literature review," Array, vol. 23, p. 100361, 2024.
[48] Z. Xue, T. Gonsalves, "Vision Based Drone Obstacle Avoidance by Deep Reinforcement Learning," AI, vol. 2, no. 3, pp. 366-380, 2021.
[49] F. AlMahamid, K. Grolinger, "Autonomous Unmanned Aerial Vehicle navigation using Reinforcement Learning: A systematic review,» Engineering Applications of Artificial Intelligence, vol. 115, p. 105321, 2022.
[50] A. K. Shakya, G. Pillai, S. Chakrabarty, "Reinforcement learning algorithms: A brief survey," Expert Systems with Applications, vol. 231, p. 120495, 2023.
[۵۱] ع. مهرابی، م. شــهبازی خجســته، ع. ا. صدری، ج. دلدارشیخی، «مسیریابی و هدایت دســته جمعی مجموعه ای از پهپاد ها در محیط های ناشناخته به منظور هــدف یابی با هوش تجمعی خودمختار،» اولین همایش ملی علوم و فناوری های نوظهور و شالوده شکن در حوزه دفاعی، ۱۴۰۳.
[52] A. P. Kalidas, C. J. Joshua, A. Q. Md, S. Basheer, S. Mohan, S. Sakri, "Deep Reinforcement Learning for Vision-Based Navigation of UAVs in Avoiding Stationary and Mobile Obstacles," Drones, vol. 7, no. 4, p. 245, 2023.
[53] M. S. Khojasteh, A. Salimi-Badr, "Autonomous Quadrotor Path Planning Through Deep Reinforcement Learning With Monocular Depth Estimation," IEEE Open Journal of Vehicular Technology, vol. 6, pp. 34-51, 2025.
[54] A. Singla, S. Padakandla, S. Bhatnagar, "Memory-Based Deep Reinforcement Learning for Obstacle Avoidance in UAV With Limited Environment Knowledge," IEEE Trans. Intell. Transport. Syst., vol. 22, no. 1, pp. 107-118, 2021.
[55] Y. Chen, N. González-Prelcic, R. W. Heath, "Collision-Free UAV Navigation with a Monocular Camera Using Deep Reinforcement Learning," 2020, pp. 1-6.
[56] C. Wang, J. Wang, Y. Shen, X. Zhang, "Autonomous Navigation of UAVs in Large-Scale Complex Environments: A Deep Reinforcement Learning Approach," IEEE Transactions on Vehicular Technology, vol. 68, no. 3, pp. 2124-2136, 2019.
[57] L. Graesser, W. L. Keng, Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional, 2019.
[58] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, "Proximal policy optimization algorithms,» arXiv preprint arXiv:1707.06347, 2017.
[59] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," arXiv preprint arXiv:1506.02438, 2015.
[60] S. Shah, D. Dey, C. Lovett, A. Kapoor, "AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles," Cham, 2018: Springer International Publishing, in Field and Service Robotics, pp. 621-635.
[61] B. Kabas, "Autonomous UAV Navigation via Deep Reinforcement Learning Using PPO," in 2022 30th Signal Processing and Communications Applications Conference (SIU), 2022, pp. 1-4.