یادگیری تقویتی عمیق برای تخصیص هوشمند منابع در شبکه‌های یکپارچه زمینی—غیرزمینی با سطح بازپیکربندی فعال(Active RIS)

محمد احمد محسن[1]، حسن رضوان[2]، محمد جزیب[3,4]، محمد اقبال[4]، محمد بلال[5]، تبنده اشرف[4]، محمد فرحان خان[6]، جن-یی پان[4]

[1]دانشکده مهندسی برق، دانشگاه استنفورد، آمریکا

[2]دانشکده مهندسی برق، دانشگاه کالیفرنیا، ریورساید، آمریکا

[3]دانشکده مهندسی برق، مؤسسه مهندسی و علوم کاربردی پاکستان، پاکستان

[4]دانشکده مهندسی ارتباطات، دانشگاه چونگ چنگ، تایوان

[5]دانشکده مهندسی برق، دانشگاه کالیفرنیا، ارواین، آمریکا

[6]دانشکده علوم کامپیوتر و فناوری اطلاعات، کالج دانشگاهی کورک، ایرلند

ایمیل نویسندگان مسئول{muahmed@stanford.edu, hrizw002@ucr.edu, iqbal.marjan@gmail.com}:

چکیده--این کار به بررسی به‌کارگیری سطوح هوشمند بازپیکربندی‌پذیر فعال (A-RIS) در شبکه‌های یکپارچه زمینی و غیرزمینی (Terrestrial and Non-Terrestrial Networks – TN-NTN) با استفاده از دسترسی چندگانه غیرمتعامد چندنقطه‌ای هماهنگ‌شده (Coordinated Multipoint Non-Orthogonal Multiple Access – CoMP-NOMA)می‌پردازد. مدل سیستم پیشنهادی شامل یک RIS مبتنی بر پهپاد است که به‌صورت هماهنگ با یک RIS زمینی عمل می‌کند و هدف آن تقویت سیگنال می‌باشد. هدف اصلی ما بیشینه‌سازی نرخ مجموع برای تمامی کاربران موجود در شبکه با استفاده از یک الگوریتم سفارشی بهینه‌سازی سیاست مجاورتی ترکیبی (Hybrid Proximal Policy Optimization – H-PPO)از طریق بهینه‌سازی مسیر حرکت پهپاد، ضرایب تخصیص توان ایستگاه پایه(Base Station – BS) ، ضریب تقویت RIS فعال و ماتریس شیفت فاز است. ما کاربران لبه سلول را در زوج‌های NOMA ادغام می‌کنیم تا بهره تنوع حاصل شود که این امر تجربه کلی کاربران لبه را بیش از پیش بهبود می‌بخشد. مقایسه‌های جامع و گسترده‌ای با شبکه‌های مجهز به RIS غیرفعال انجام شده است تا کارایی برتر RIS فعال از نظر بهره‌وری انرژی، احتمال قطعی و نرخ مجموع شبکه به‌طور واضح نشان داده شود.

واژگان کلیدی-- RIS فعال، یادگیری تقویتی عمیق، NOMA، تخصیص منابع، TN-NTN

I. مقدمه

گسترش روزافزون دستگاه‌های بی‌سیم منجر به افزایش چشمگیر تقاضا برای منابع ارتباطی شده است. طرح‌های سنتی دسترسی چندگانه متعامد(Orthogonal Multiple Access – OMA) ، که در آن‌ها منابع زمانی–فرکانسی اختصاصی به هر کاربر تخصیص داده می‌شود، به‌تدریج در حال تبدیل شدن به روش‌هایی ناکارآمد هستند. دسترسی چندگانه غیرمتعامد (Non-Orthogonal Multiple Access (NOMA – به‌عنوان یک فناوری نویدبخش مطرح شده است که با بهره‌گیری از کدگذاری برهم‌نهی (Superposition Coding)و حذف تداخل متوالی – (Successive Interference Cancellation (SIC)، امکان اشتراک‌گذاری منابع زمانی–فرکانسی یکسان را برای چندین کاربر فراهم می‌کند. این ویژگی منجر به افزایش بهره‌وری طیفی سیستم می‌شود [1]، [2]. با تخصیص توان به کاربران بر اساس شرایط کانال، NOMA قادر است عدالت را در سیستم بهبود داده و کاربران بیشتری را در یک طیف فرکانسی مشخص پشتیبانی کند .در محیط‌های چندسلولی، کاربران لبه سلول اغلب از تداخل شدید بین‌سلولی رنج می‌برند، زیرا این کاربران در نزدیکی مرزهای سلول‌ها قرار دارند و بیشتر در معرض تداخل ناشی از سلول‌های مجاور هستند .به‌منظور کاهش این مشکل، از انتقال چندنقطه‌ای هماهنگ‌شده (CoMP) استفاده می‌شود که امکان آشکارسازی مشترک و انتقال مشترک هماهنگ‌شده میان چندین ایستگاه پایه را فراهم می‌سازد؛ امری که می‌تواند عملکرد کاربران لبه سلول را به‌طور قابل‌توجهی بهبود دهد[3] . RIS زمینی معمولاً در مناطق شهری با تراکم بالا که تقاضای کاربران پایدار است، مستقر می‌شود، در حالی که RIS فعال (A-RIS) برای محیط‌های پویایی به کار می‌رود که در آن‌ها تقاضای کاربران به‌سرعت تغییر می‌کند [4]، [5] . علاوه بر این، در محیط‌های شهری موانعی برای پرواز پهپاد وجود دارد که نواحی بلااستفاده‌ای در مدل ایجاد می‌کنند و به‌عنوان مناطق ممنوعه پرواز برای پهپاد در نظر گرفته می‌شوند. بنابراین، بهینه‌سازی مسیر حرکت پهپاد به‌منظور بهبود نرخ مجموع و تضمین ایمنی پهپاد امری ضروری است. برخی از پژوهش‌ها از پهپاد به‌عنوان ایستگاه پایه استفاده می‌کنند که در مقایسه با RIS انعطاف‌پذیری کمتری دارد [6]، چرا که RISها می‌توانند به‌صورت پویا بازپیکربندی شوند تا ویژگی‌های بازتابی خود را بر اساس شرایط متغیر کانال تنظیم کرده و حتی پرتوها را به سمت کاربران هدایت کنند تا ظرفیت شبکه افزایش یابد. به‌طور معمول، سیستم‌های ارتباطی مبتنی بر پهپاد ازRIS های غیرفعال استفاده می‌کنند و ماتریس شیفت فاز آن‌ها را به‌منظور بهبود توان عملیاتی سیستم بهینه‌سازی می‌نمایند. با این حال، استفاده از RIS فعال و بهینه‌سازی ماتریس ضریب تقویت آن تأثیر قابل‌توجهی بر نرخ مجموع کلی محیط خواهد داشت، زیرا مقایسه ارائه‌شده در [7] نشان می‌دهد که RIS فعال بر اثر تضعیف ضربی موجود در RISغیرفعال غلبه کرده و در نهایت بهره نرخ مجموع بالاتری را فراهم می‌کند. در حالی که یادگیری تقویتی عمیق (Deep Reinforcement Learning – DRL)به‌طور فزاینده‌ای برای بهینه‌سازی سیستم‌های شامل RIS و پهپاد مورد استفاده قرار گرفته است، پژوهش‌های موجود عمدتاً بر فضاهای کنش پیوسته یا گسسته تمرکز داشته‌اند.[8] برای دستیابی به عملکرد بهینه، یک رویکرد ترکیبی که هر دو فضای کنش پیوسته و گسسته را در بر بگیرد، ضروری است. الگوریتم‌های DRL که قادر باشند به‌طور هم‌زمان از هر دو نوع فضای کنش استفاده کنند، می‌توانند راه‌حل‌های بسیار امیدوارکننده‌ای برای مسائل بهینه‌سازی ارائه دهند. علاوه بر این، بهینه‌سازی ماتریس تقویت RIS فعال با استفاده از DRL تاکنون مورد بررسی قرار نگرفته است. با در نظر گرفتن این شکاف در ادبیات پژوهشی به‌عنوان انگیزه اصلی، مشارکت‌های این مقاله به‌صورت سه‌گانه ارائه می‌شوند. در این کار، یک عامل یادگیری تقویتی عمیق ترکیبی با نام H-PPO برای بهینه‌سازی شیفت‌های فاز، مسیر حرکت پهپاد، ضرایب تخصیص توان NOMA و توان ارسالی ایستگاه پایه به‌منظور دستیابی به بیشینه نرخ مجموع تحت شرایط SNR بهینه به کار گرفته شده است.

II. مدل سیستم و صورت‌بندی مسئله

A. توصیف سیستم

ما یک شبکه انتقال پایین‌دست (downlink) تک‌ورودی تک‌خروجی (SISO) مبتنی بر CoMP-NOMAبا کمک RIS فعال و استقرار پهپاد را در نظر می‌گیریم. شبکه در قالب M ناحیه دایره‌ای توزیع شده است که در مرکز هر ناحیه، یک ایستگاه پایه تک‌آنتنه (BS) قرار دارد، به‌طوری‌که $m \in M = \{1,2,\dots,M\}$ هر ایستگاه پایه BS_m در ناحیه m از NOMA دوکاربره برای فراهم‌سازی کانال پایین‌دست برای کاربران مرکزی و کاربران لبه استفاده می‌کند. کاربران مرکزی درون شعاع ناحیه قرار دارند و توسط BS_m سرویس‌دهی می‌شوند و با U_c^m نمایش داده می‌شوند، که$c \in C = \{1,2,\dots,C\}$ نمایانگر مجموعه کاربران مرکزی است. در مقابل، کاربران لبه خارج از شعاع ناحیه قرار دارند و با U_e^m نمایش داده می‌شوند که توسط BS_m سرویس می‌گیرند، به‌طوری‌که $e \in \varepsilon = \{1,2,\dots,E\}$ نمایانگر مجموعه کاربران لبه است. تمام کاربران موجود در سیستم را می‌توان به‌صورت $U = C \cup \varepsilon$ نشان داد، به‌طوری‌که هر کاربر $u \in U$ یا یک کاربر مرکزی است یا یک کاربر لبه. کل زمان پرواز پهپاد به t بازه زمانی تقسیم می‌شود، به‌طوری‌که $t \in \mathcal{T} = \{1,2,\dots,t_s\}$، که در آن t_sکل زمان پرواز پهپاد است. علاوه بر این، پهپاد اجازه پرواز در نواحی ممنوعه موجود در مدل سیستم را ندارد که این نواحی به‌صورت دیسک‌های دایره‌ای با شعاع d_{min} و مرکزیت موانع مختلف مدل‌سازی می‌شوند، به‌طوری‌که $o \in \mathcal{O} = \{1,2,\dots,O\}$. مدل سیستم شامل دو RIS فعال است: یکی روی یک ساختمان و در فاصله‌ای برابر از تمامی نواحی شبکه با نماد نصب شده است و دیگری روی یک پهپاد متحرک با نماد قرار دارد که موقعیت خود را برای فراهم‌سازی بیشینه نرخ مجموع کاربران شبکه تغییر می‌دهد. هر دو RIS به تمامی کاربران شبکه سرویس می‌دهند و توسط یک ریزکنترل‌گر کنترل می‌شوند تا شیفت‌های فاز را تغییر دهند. RIS فعال با نماد $R \in R_G, R_U$ در مدل سیستم به K عنصر بازتابنده مجهز است، به‌طوری‌که $k \in \{1,2,\dots,K\}$. سیگنال‌هایی که چندین بار توسطRIS ها بازتاب می‌شوند، به دلیل تلفات شدید مسیر، دارای توان ناچیز فرض می‌شوند؛ همان‌گونه که در محیط‌های چند RIS-در [9] در نظر گرفته شده است و این پژوهش نیز همین فرض را اتخاذ می‌کند .به‌طور خاص، برای $u \in U, \forall m \in \mathcal{M}$ ، $o \in \mathcal{O}$ ، موقعیت ایستگاه پایه BS_m، کاربر U_u و مانع O_o به‌ترتیب به‌صورت زیر نمایش داده می‌شوند: $p_m = \left(x_m, y_m, H_B\right)$، $p_o = (x_o, y_o, H_o)$,$p_u = (x_u, y_u, 0)$ که در آن H_B و H_o به‌ترتیب ارتفاع ایستگاه‌های پایه و موانع هستند .موقعیت RIS هوایی در بازه زمانی به‌صورت $p_{R_U}[t] = (x_{R_U}[t], y_{R_U}[t], H_{R_U})$ نمایش داده می‌شود و این RIS بر فراز یک ناحیه مشخص A شناور است .همان‌طور که پیش‌تر ذکر شد، پهپاد متحرک بوده و موقعیت افقی خود را در صفحه تغییر می‌دهد، در حالی که ارتفاع آن H_{R_U} ثابت باقی می‌ماند.

B. مدل کانال

مدل در نظر گرفته‌شده شامل مسیرهای دید مستقیم (LoS) و بدون دید مستقیم (NLoS) است؛ ازاین‌رو، از مدل‌های محوشدگی رایلی (Rayleigh) و رایسین (Rician) استفاده می‌شود. به دلیل پراکندگی موجود در محیط، از مدل کانال محوشدگی رایلی برای شبیه‌سازی کانال‌های بین ایستگاه‌های پایه و کاربران استفاده می‌شود.

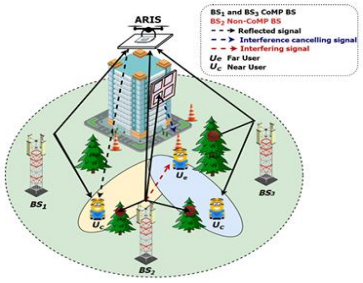


Fig. 1: RIS-assisted TN-NTN with coordinated NOMA pairing.

کانال بین ایستگاه پایه BS_m و کاربر U_u با $h_{m,u}$نمایش داده می‌شود و برای بازه زمانی t به‌صورت زیر بیان می‌گردد:

$$h_{m,u}[t] = \sqrt{\frac{\rho_0}{PL(d_{m,u})}}\, v_{m,u}[t], \quad (1)$$

که در آن ρ_0 تلفات مسیر مرجع در فاصله ۱ متر است و $PL(d_{m,u})$ تلفات مسیر مقیاس بزرگ بوده که به‌صورت $PL(d_{m,u}) = (d_{m,u})^{\alpha}$مدل‌سازی می‌شود؛ در اینجا α توان تلفات مسیر و $d_{m,u}$ فاصله بین BS_mو کاربر U_u است. همچنین، ضریب محوشدگی رایلی مقیاس کوچک با $v_{m,u}\sim R(1)$ مدل‌سازی می‌شود .از آنجا که بین ایستگاه پایه و RIS نصب‌شده روی پیهاد و همچنین RIS نصب‌شده روی ساختمان، دید مستقیم وجود دارد، این کانال با استفاده از محوشدگی رایسین شبیه‌سازی می‌شود .کانال بین BS_m و RIS با نماد $h_{m,R}$ نمایش داده شده و برای بازه زمانی t به‌صورت زیر بیان می‌شود

$$h_{m,u}[t] = \sqrt{\frac{\rho_0}{PL(d_{m,u})}}\left(\sqrt{\frac{K}{1+K}}\, g_{m,R}^{Los} + \sqrt{\frac{1}{1+K}}\, g_{m,R}^{NLoS}\right), \quad (2)$$

که در آن K عامل رایسین است. بردار کانال دید مستقیم $g_{m,R}^{Los} \in \mathbb{C}^{K\times 1}$ به‌صورت زیر تعریف می‌شود:

$$g_m^{Los} = [1, \dots, e^{j(k-1)\pi \sin(\omega)}, \dots, e^{j(K-1)\pi \sin(\omega)}]^T,$$

که در آن ω زاویه ورود مؤلفه دید مستقیم به RIS را نشان می‌دهد .مولفه بدون دید مستقیم $g_{m,R}^{NLoS} \in \mathbb{C}^{K\times 1}$ از محوشدگی رایلی پیروی می‌کند، همان‌گونه که در [3] توصیف شده است .در این پژوهش، برای جلوگیری از سربارهای غیرضروری، اطلاعات حالت کانال کامل (Perfect CSI) فرض شده است.

C. پیکربندی RIS فعال

RIS فعال موجود در شبکه، سیگنال ورودی را تقویت می‌کند که این امر ناشی از استفاده از تقویت‌کننده‌های نوع بازتابی در عناصر RIS فعال است که به یک منبع تغذیه متصل هستند. سیگنال خروجی RIS فعال با K عنصر بازتابنده به‌صورت زیر بیان می‌شود:

$$y[t] = \underbrace{P[t]\Theta[t]x[t]}_{\text{Desired signal}} + \underbrace{P[t]\Theta[t]v}_{\text{Dynamic noise}} + \underbrace{n_s}_{\text{Static noise}}, \quad (3)$$

که در $P[t] = diag(p_1[t], \dots, p_K[t]) \in \mathbb{R}^{K\times K}$ ماتریس تقویت RIS فعال در بازه زمانی t است .هر عنصر RIS فعال به‌صورت $1 \le p_k \le S$ مدل‌سازی می‌شود، که در آن S بیشینه میزان تقویتی است که هر عنصر می‌تواند فراهم کند. همچنین، $x[t] \in \mathrm{C}^{k}$ سیگنال ورودی RIS ، $y[t] \in \mathrm{C}^{k}$ سیگنال خروجی RIS و $\Theta[t] = diag(a_1 e^{j\theta_1[t]}, a_2 e^{j\theta_2[t]}, \dots, a_K e^{j\theta_K[t]}) \in \mathbb{C}^{K\times 1}$ ماتریس شیفت فاز RIS در بازه زمانی t است، که در آن $a_k \in (0,1]$ ضریب دامنه و $\theta_k \in [-\pi, \pi]$ شیفت فاز عنصر K-ام RIS را نشان می‌دهد .عملگر diag(0) به عملیات قطری‌سازی اشاره دارد . عناصر RIS فعال برای تقویت سیگنال‌های بازتابی، توان اضافی مصرف کرده و نویز حرارتی تولید می‌کنند که به دو بخش نویز پویا و نویز ایستا تقسیم می‌شود .در این کار، تنها نویز پویا در نظر گرفته شده و نویز ایستا نادیده گرفته می‌شود، زیرا مطالعه [7] نشان می‌دهد که نویز ایستا نسبت به نویز پویا به‌طور قابل‌توجهی کوچک‌تر بوده و می‌توان از آن صرف‌نظر کرد .نویز پویا با متغیر v نمایش داده شده و به‌صورت زیر مدل می‌شود : $v\sim \mathcal{CN}(0_K, \sigma^2 I_K)$که در آن $\mathcal{CN}(\mu, \Sigma)$ توزیع گاوسی مختلط چندمتغیره با میانگین μ و کوواریانس Σ را نشان می‌دهد .همچنین I_K ماتریس همانی $K \times K$ و 0_K بردار صفر $K \times 1$ است.

D. مدل NOMA

تمام ایستگاه‌های پایه یک کاربر مرکزی و یک کاربر لبه را سرویس‌دهی می‌کنند؛ بنابراین، مطابق با سنت NOMA دوکاربره، توان ایستگاه پایه بین دو کاربر توزیع می‌شود، به‌طوری‌که کاربر لبه توان بیشتری دریافت می‌کند [10]، [11]. سیگنال ارسالی توسط ایستگاه پایه BS_m به‌صورت زیر بیان می‌شود ، $x_m[t] = \sqrt{(1-\lambda_m)p_t}x_{m,c}[t] + \sqrt{\lambda_m p_t}x_{m,e}[t]$ که در آن $x_{m,c}[t]$ و $x_{m,e}[t]$ به‌ترتیب سیگنال دریافتی برای کاربر مرکزی U_c و کاربر لبه U_e هستند. تمام ایستگاه‌های پایه در سیستم دارای توان ارسال یکسان p_t هستند و ضریب تخصیص توان کاربر لبه با λ_m نمایش داده می‌شود. ضریب تخصیص توان λ_m باید قید زیر را ارضا کند،$\lambda_m < 1$ و $0.5 < \lambda_m$ زیرا کاربر لبه فاصله بیشتری از ایستگاه پایه دارد و نیازمند تخصیص توان بالاتری است[11] .

هر کاربر مرکزی U_c یک سیگنال مستقیم از ایستگاه پایه، یک لینک غیرمستقیم بازتابی ازRIS ، تداخل ناشی از ایستگاه پایه دیگر که کاربر سلول مجاور را سرویس می‌دهد، و همچنین نویز ذاتی RISفعال را دریافت می‌کند. در نهایت، سیگنال دریافتی توسط کاربر مرکزی به‌صورت زیر بیان می‌شود:

$$y_c[t] = \underbrace{H_{m,c}[t]p_t x_m[t]}_{\text{desired signal}} + \underbrace{\sum_{j\neq m}^{M} H_{j,c}[t]p_t x_j[t]}_{\text{inter-user interference}} + \underbrace{h_{R,c}[t]P[t]\Theta[t]v}_{\text{active RIS noise}} + n_u, \quad (4)$$

که در آن n_uنویز سفید گاوسی افزایشی برای کاربرu بوده و به‌صورت $n_u\sim\mathcal{N}(0,\sigma^2)$ مدل‌سازی می‌شود. کانال بین BS_m و کاربر مرکزی U_c به‌صورت زیر تعریف می‌شود، $H_{m,c}[t] = h_{m,c}[t] + h_{R,c}[t]P[t]\Theta[t]h_{m,R}[t]$ که شامل لینک مستقیم و لینک غیرمستقیم از طریق RIS است .تداخل بین‌کاربری (ICI) زمانی رخ می‌دهد که U_cسیگنال پخش‌شده از ایستگاه پایه BS_j را دریافت کند [12]. پس از به‌کارگیری حذف تداخل متوالی(SIC) ، کاربر مرکزی ابتدا سیگنال مربوط به کاربر لبه را آشکارسازی کرده و آن را از سیگنال دریافتی حذف می‌کند تا بتواند سیگنال خود را رمزگشایی نماید. بنابراین، نرخ دست‌یافتنی برای کاربر مرکزی جهت آشکارسازی سیگنال کاربر لبه به‌صورت زیر است:

$$\mathcal{R}_{c\to e}[t] = \log\left(1 + \frac{\lambda_m\gamma_m}{(1-\lambda_m)\gamma_m + 1}\right), \quad (5)$$

که در آن $\gamma_m = \frac{p_t|h_{m,c}[t]|^2}{\psi}$ و $\psi = p_t|h_{j,c}[t]|^2 + \sigma^2$ در مقابل، نرخ داده‌ای که کاربر مرکزی برای آشکارسازی سیگنال خود به‌دست می‌آورد به‌صورت زیر بیان می‌شود

$$\mathcal{R}_c[t] = \log\left(1 + (1-\lambda_m)\gamma_m\right). \quad (6)$$

سیگنال دریافتی توسط کاربر لبه از طریق BS_m و RIS به‌صورت زیر بیان می‌شود

$$y_e = (h_{m,e}[t] + h_{R,e}^H[t]P[t]\Theta[t]h_{m,R}[t])x_m[t]p_t + (h_{j,e}[t] + h_{R,e}^H[t]P[t]\Theta[t]h_{j,R}[t])x_j[t]p_t + n_u, \quad (7)$$

که در آن $h_{R,e}^H$ ترانهاده مزدوج (Hermitian) کانال بین R RIS و کاربر لبه e را نشان می‌دهد .نرخ داده قابل دستیابی برای کاربر لبه به‌صورت زیر تعریف می‌شود:

$$\mathcal{R}_e[t] = \log\left(1 + \frac{\lambda_m\gamma_m + \lambda_j\gamma_j}{(1-\lambda_m)\gamma_j + (1-\lambda_j)\gamma_j + 1}\right) \quad (8)$$

پس از به‌دست آوردن نرخ مجموع برای کاربر مرکزی در رابطه (6) و کاربر لبه در رابطه (8)، نرخ مجموع کل سیستم به‌صورت زیر محاسبه می‌شود:

$$R_{\text{total}}[t] = \sum_{c\in\mathcal{C}} R_c[t] + \sum_{e\in\mathcal{E}} R_e[t]. \quad (9)$$

E. بهره‌وری انرژی

بهره‌وری انرژی سیستم پیشنهادی NOMA-CoMP مبتنی بر RIS فعال، موازنه میان عملکرد شبکه و هزینه‌های انرژی را برجسته می‌کند که به‌صورت زیر تعریف می‌شود:

$$\eta_E = \frac{R_{\text{total}}[t]}{p_{\text{total}}}. \tag{10}$$

که در آن $R_{total}[t]$ نرخ مجموع سیستم در بازه زمانی t بوده و p_{total} توان کل مصرفی شبکه را نشان می‌دهد .توان کل مصرفی شبکه شامل توان مصرفی ایستگاه‌های پایه، RISفعال و پهپاد است .با توجه به اینکه نرخ مجموع کل سیستم در رابطه (9) بیان شده است، توان کل مصرفی شبکه را می‌توان به‌صورت زیر نوشت:

$$p_{\text{total}} = \underbrace{p_{\text{BS}}}_{\text{BS power}} + \underbrace{\sum_{i=1}^{2}\sum_{k=1}^{K}\left(\eta_{R_i}|P_{R_ik}\Theta_{R_ik}x_{R_ik}|^2 + P_{\text{c},R_ik}\right)}_{\text{RIS power}} + \underbrace{P_{\text{UAV}}}_{\text{UAV power}}, \tag{11}$$

که در آن p_{BS} نشان‌دهنده توان ایستا و توان ارسال ایستگاه پایه است؛ توان مصرفی RIS شامل η_Rبه‌عنوان بازده تقویت RIS R_i ، P_{R_ik} توان سیگنال ورودی به عنصر -ام RIS R_i ،شیفت فاز اعمال‌شده توسط عنصر -ام RIS R_i ، x_{R_ik} سیگنال ورودی به عنصر -ام، و P_{c,R_ik} توان مصرفی مدار عنصر -ام RIS R_i می‌باشد؛ در نهایت، P_{UAV} توان مصرفی پهپاد برای شناوری و حرکت آن را نشان می‌دهد.

F. صورت‌بندی مسئله

در این پژوهش، هدف اصلی ما بیشینه‌سازی نرخ مجموع حاصل‌شده در طول بازه‌های زمانی t است . برای دستیابی به این هدف، چهار متغیر کنترلی کلیدی به‌صورت هم‌زمان بهینه‌سازی می‌شوند . مسیر حرکت پهپاد $\hat{\mathcal{T}} \triangleq \{p_{Ru}[t], \forall t$ و $\Theta \triangleq \{\Theta\ [t], \forall t$ و $\Lambda \triangleq \{\lambda_m, \forall m\}$ و $\{p_k[t], \forall k\}$ مسئله P بهینه‌سازی به‌صورت ریاضی به شکل زیر صورت‌بندی می‌شود:

$$\max_{\mathbf{P},\boldsymbol{\Theta},\boldsymbol{\Lambda},\hat{\mathbf{T}}} \sum_{t\in\mathcal{T}} R_{\text{sum}}[t], \tag{10a}$$

به‌طوری‌که قیود زیر برقرار باشند:

$$x_R[t], y_R[t] \in A, \quad \forall t \in \mathcal{T}, \tag{10b}$$
$$\|\mathbf{p}_R[t] - \mathbf{p}_o\| \geq d_{\min}, \quad \forall o \in \mathcal{O}, t \in \mathcal{T}, \tag{10c}$$
$$\theta_k[t] \in [-\pi, \pi), \quad \forall k \in K, t \in \mathcal{T}, \tag{10d}$$
$$R_c[t] \geq R_c^{\min}, \quad \forall m \in M, t \in \mathcal{T}, \tag{10e}$$
$$R_e[t] \geq R_e^{\min}, \quad \forall e \in \mathcal{E}, t \in \mathcal{T}, \tag{10f}$$
$$\lambda_m \in (0.5, 1), \quad \forall m \in M, t \in \mathcal{T}, \tag{10g}$$
$$p_k[t] \in [1, \mathcal{S}], \quad \forall k \in K, t \in \mathcal{T}, \tag{10h}$$

III. راهکار پیشنهادی مبتنی بر یادگیری تقویتی عمیق

A. صورت‌بندیMDP

فرآیند تصمیم‌گیری مارکوف (MDP) با استفاده از پنج‌تایی تعریف می‌شود، که در آن و به‌ترتیب فضای حالت و فضای کنش را نشان می‌دهند .نماد پاداشی است که عامل در ازای اعمال خود دریافت می‌کند و عامل تنزیل (discount factor) برای ایجاد توازن میان پاداش‌های آنی و آینده است .در نهایت، احتمال انتقال حالت را نشان می‌دهد، یعنی احتمال انتقال از یک حالت به حالت دیگر در نتیجه انجام یک کنش .در هر بازه زمانی ، عامل حالت جاری را مشاهده می‌کند، بر اساس سیاست خود یک کنش را انتخاب می‌نماید و به حالت جدید منتقل می‌شود، در حالی که پاداش متناظر را دریافت می‌کند.

B. فضای حالت

فضای حالت در بازه زمانی شامل موقعیت فعلی پهپاد، ضرایب تخصیص توان ایستگاه پایه، ماتریس تقویت RIS فعال و نرخ‌های مجموع قابل دستیابی است که به‌ترتیب با $R[t] = \mathcal{Y}v[t], \varsigma, \varrho[t]$ نمایش داده می‌شوند. در نهایت، فضای حالت به‌صورت زیر تعریف می‌شود:

$$s_t = \{\varrho[t], \zeta, \vartheta[t], \mathbf{R}[t]\}. \tag{12}$$

بعد فضای حالت به‌صورت زیر بیان می‌شود $\dim(S_t) = 2 + M + K^2 + 2$.

C. فضای کنش

فضای کنش MDP شامل حرکت افقی پهپاد در صفحه ، ضریب تقویت و شیفت‌های فاز RIS فعال، و ضرایب تخصیص توان ایستگاه‌های پایه است.

به‌طور خاص، فضای کنش در بازه زمانی شامل کنش‌های حرکتی پهپاد است که به‌ترتیب بیانگر حرکت به چپ، راست، پایین، بالا و حالت شناوری (hover)می‌باشند. شیفت‌های فاز با،ضرایب تخصیص توان با، و ماتریس تقویت RIS فعال با نمایش داده می‌شوند. بنابراین، فضای کنش به‌صورت زیر تعریف می‌شود:

$$a_t = \{a_R[t], a_\theta[t], a_\Lambda, a_p[t]\} \tag{13}$$

بعد فضای کنش برابر است با $\dim(a_t) = 2 + k + M + K^2$.

Algorithm 1 H-PPO Active RIS-Based Energy Optimization

Initialize: Parameters $\alpha_d, \alpha_c, \zeta, \Psi$
for iteration $l = 1, \ldots, L$ **do**
  Receive initial state s_0
  for time step $t = 0, \ldots, T$ **do**
    Choose continuous actions a_p and a_c using $\pi_{\alpha_c}(s_t)$
    for each RIS element n **do**
      Energy optimization $e_n = g(\psi_n, a_p, a_c)$
      Update the gain for the k-th RIS element based on channel conditions
    end for
    Execute actions $a_t = \{a_R, \Psi, a_p, a_c\}$
    Observe reward r_t based on achieved sum rate and outage probability
    Store the transition (s_t, a_t, r_t, s_{t+1}) in experience replay buffer
    Compute the advantage function estimate $\hat{A}_t$
  end for
end for
for epoch $m = 1, \ldots, K$ **do**
  Sample mini-batch experiences E from the replay buffer
  Objective computation using clipped functions for both discrete and continuous actions: $L_d^{\text{CLIP}}(\theta_d)$ and $L_c^{\text{CLIP}}(\theta_c)$
  Optimize policy parameters α_d and α_c
end for
Obtain previous policy parameters $\alpha_d^{\text{old}} \leftarrow \alpha_d$ and $\alpha_c^{\text{old}} \leftarrow \alpha_c$
Clear

D. تابع پاداش

تابع پاداش به‌گونه‌ای طراحی شده است که ضمن بیشینه‌سازی نرخ مجموع، ایمنی پهپاد و برآورده شدن الزامات کیفیت سرویس (QoS) را نیز تضمین کند. در صورتی که پهپاد از محدوده مجاز خارج شود(Out of Bounds) ، نقض قیود با اعمال جریمه در تابع پاداش لحاظ می‌شود. تابع پاداش به‌صورت زیر تعریف می‌شود:

$$R(s_t, a_t) = R_{\text{sum}}[t] + \xi_{\text{dist}}\left(\frac{C}{d_{\mathbf{R}_U,\mathbf{U}}[t]}\right)\zeta - \xi_{\text{OOB}}\cdot\mathbb{I}(\text{OOB}), \tag{14}$$

که در آن تابع شاخصی است که برای تشویق پهپاد به نزدیک ماندن به کاربران به کار می‌رود. همچنین، جریمه‌ای است که در صورت خروج RIS هوایی از محدوده مجاز به عامل اعمال می‌شود، که توسط تابع شاخص مشخص می‌گردد. در نهایت، و به‌عنوان ضرایب ثابت در این رابطه در نظر گرفته می‌شوند.

E. الگوریتم H-PPO

برای مدیریت فضای کنش ترکیبی در شبکه، از الگوریتم بهینه‌سازی سیاست مجاورتی ترکیبی (Hybrid Proximal Policy Optimization – H-PPO)استفاده می‌شود. این الگوریتم شامل کنش‌های گسسته و کنش‌های پیوسته، و است. تابع ارزش برای به‌دست آوردن تخمین تابع مزیت با واریانس کاهش‌یافته جهت بهینه‌سازی سیاست مورد استفاده قرار می‌گیرد. بر اساس جزئیات پیاده‌سازی ارائه‌شده در [13]، سیاست برای بازه زمانی اجرا شده و به‌صورت زیر محاسبه می‌شود:

$$\hat{A}_t = \sum_{k=0}^{\bar{T}-1} \mathbb{D}^k r_{t+k} + \mathbb{D}^{\bar{T}}\hat{V}(s_{t+\bar{T}}) - V(s_t), \tag{15}$$

که در آن به‌طور قابل‌توجهی کوچک‌تر از طول اپیزود است. برای تولید سیاست تصادفی کنش‌های گسسته، شبکه بازیگر (Actor) گسسته تعداد لاجیت خروجی تولید می‌کند که سپس از تابع softmaxعبور داده می‌شوند تا توزیع احتمال روی کنش‌های گسسته حاصل شود. در مقابل، شبکه

بازیگر پیوسته، کنش‌های پیوسته و را با نمونه‌برداری از توزیع‌های گاوسی که با میانگین و انحراف معیار خروجی شبکه پارامتردهی شده‌اند، تولید می‌کند؛ مطابق با سیاست تصادفی هر دو سیاست و به‌طور مستقل و با استفاده از توابع هدف کلیپ‌شده مربوطه بهینه‌سازی می‌شوند. تابع هدف برای کنش‌های گسسته به‌صورت زیر تعریف می‌شود:

$$L_d^{\text{CLIP}}(\theta_d) = \hat{\mathbb{E}}_t \left[\min \left(r_t^d(\theta_d)\hat{A}_t, \aleph(r_t^d, \theta_d, \epsilon)\hat{A}_t \right) \right] \qquad (16)$$

که در آن نسبت نمونه‌برداری اهمیت است و پارامتر کلیپ‌سازی می‌باشد. اهداف بهینه‌سازی سیاست‌های گسسته و پیوسته از یکدیگر جدا هستند؛ بدین معنا که و به‌عنوان توزیع‌های مستقل در فرآیند بهینه‌سازی سیاست در نظر گرفته می‌شوند، نه یک توزیع مشترک شامل هر دو فضای کنش. خلاصه الگوریتم H-PPO در الگوریتم ۱ ارائه شده است.

IV. نتایج شبیه‌سازی

A. تنظیمات شبیه‌سازی

به‌منظور ارزیابی اثربخشی چارچوب پیشنهادی RIS فعال در یک شبکهCoMP-NOMA ، یک نمایش مجازی از یک محیط شهری ایجاد می‌شود که شامل ایستگاه پایه c=3، کاربر k=3، و مانع L=0می‌باشد .موقعیت اولیه پهپاد برابر با m (40 , 0 , -5) و [20, - , -5], [30, 30, 20], [-30, 30, 20] [30, 20] با این حال، هر ایستگاه پایه با یک سطح توان منحصربه‌فرد عمل می‌کند که با برای هر ایستگاه پایه نمایش داده می‌شود، به‌طوری‌که سایر اجزای سیستم، شامل کاربران، به‌صورت تصادفی در محیط توزیع شده‌اند. شبکه در فرکانس حامل f_c=2.4 GHZ و B=10 MHZتوان نویز به‌صورت زیر تعریف می‌شود $\sigma^2 = 174 + \log_{10}(B)\ dBm$ برای در نظر گرفتن اثرات تضعیف، توان‌های تلفات مسیر به‌صورت زیر انتخاب شده‌اند $\beta_s = 3.7, \beta_k = 3.3, \beta_o = 3.0, \beta_n = 2.2$ نتایج ارائه‌شده، میانگین‌گیری‌شده بر روی ۱۰۰۰ تحقق مستقل از موقعیت کاربران هستند تا معیارهای عملکرد قابل‌اعتمادی به دست آید.

B. نتایج عددی

در شکل 2(a)، رابطه بین توان ارسال ایستگاه پایه، که در بازه قرار دارد، و نرخ مجموع مورد بررسی قرار گرفته است .در این تحلیل، مقایسه‌ای بین پیکربندی‌های مختلف H-PPO-RIS انجام شده است که به‌طور مشخص بهبود عملکرد RIS فعال را برجسته می‌سازد .در مقایسه با پیکربندی H-PPO P-RIS، پیکربندی H-PPO A-RIS در مقدار به نرخbps/Hz 7 دست می‌یابد و عملکرد بهتری نسبت به RIS غیرفعال ارائه می‌دهد. نرخ تجمعی برای پیکربندی‌های H-PPO P-RIS و H-PPO A-RIS به‌ترتیب به bps/Hz 15و bps/Hz 20 می‌رسد، زمانی که dBm 25 باشد .این امر نشان‌دهنده بهبود ٪۳۳ در نرخ مجموع است که توانایی تقویت RIS فعال، در ترکیب باNOMA ، را برجسته می‌سازد و عملکردی بهتر از پیکربندی‌های غیرفعال ارائه می‌دهد .بررسی ما همچنین به کارایی الگوریتم یادگیری برای بهینه‌سازی انرژی گسترش یافته است .در شکل 2(b)، روند تغییرات پاداش در طول چندین گام آموزشی نشان داده شده است .پیکربندی H-PPO-NOMA رشد تندتری را از مقدار0.2 تا 1.0 نشان می‌دهد که بیانگر کارایی بالاتر یادگیری و مدیریت بهینه‌تر انرژی در NOMAاست. علاوه بر این، احتمال قطعی (Outage Probability) پیکربندی‌های مختلف نیز مورد ارزیابی قرار گرفته است. در شکل 2(c)، برای مقادیر p_t در بازه dBm 30- تا dBm 10-احتمال قطعی تمامی پیکربندی‌ها برای بدترین نرخ کاربران تقریباً یکسان باقی می‌ماند .پیکربندی H-PPO A-RIS NOMA در مقدار کمترین احتمال قطعی برابر با 0.20 را نشان می‌دهد، که بیانگر افزایش نرخ بدترین کاربر تا می‌باشد و از تمامی پیکربندی‌های دیگر عملکرد بهتری دارد. در شکل 3(a)، رابطه بین بهره‌وری انرژی و بهره‌وری طیفی برای پیکربندی‌های مختلف نشان داده شده است. پیکربندی H-PPO A-RIS NOMA بالاترین عملکرد را ارائه می‌دهد و به حداکثر بهره‌وری انرژی bit/J 11,500 و bits/s/Hz 24این بهبود ناشی از توانایی RIS فعال در تنظیم پویای بازتاب‌ها بوده و با راهبرد NOMAتقویت شده است، که در مجموع ٪۲۰ بهبود نسبت به پیکربندی‌های RIS غیرفعال را نشان می‌دهد .در شکل 3(b)، نرخ مجموع برای پیکربندی‌های مختلف با افزایش تعداد عناصر RIS بررسی شده است .تعداد عناصر K RIS از30 تا 200 افزایش می‌یابد که نشان می‌دهد افزودن عناصر بیشتر بهRIS ، توانایی سیستم را در بهینه‌سازی بازتاب‌ها افزایش می‌دهد .در تمامی مقادیر K ، پیکربندی H-PPO A-RIS NOM بیشترین مقدار R_kرا نشان می‌دهد که مزیت استفاده از RIS فعال همراه با NOMAرا برجسته می‌کند .در نهایت، شکل 3 (c)رابطه بین توان ارسال هر ایستگاه پایه و نرخ مجموع را برای پیکربندی‌های مختلفRIS ، در حالت‌های با عدالت و بدون عدالت، نشان می‌دهد .با افزایش از dBm 5- تا dBm 25 نرخ مجموع برای تمامی پیکربندی‌ها افزایش می‌یابد. در حالتی که توزیع توان به‌صورت یکنواخت بین کاربران مرکزی و لبه انجام شود، اعمال عدالت منجر به کاهش جزئی نرخ مجموع می‌شود. با این حال، پیکربندی H-PPO A-RIS NOMA در حالت همراه با عدالت، همچنان بهترین عملکرد را در میان تمامی پیکربندی‌ها ارائه می‌دهد.

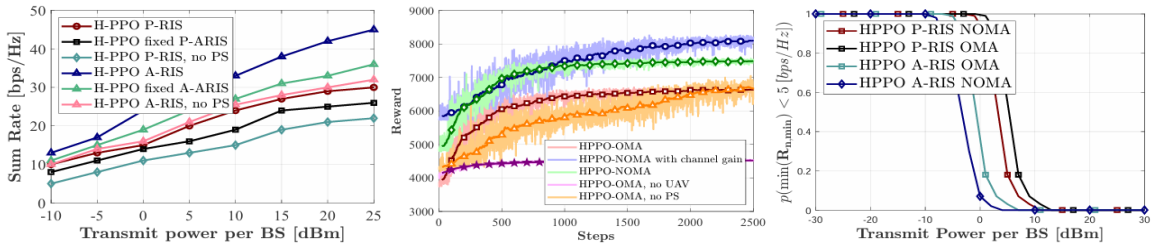
V. نتیجه‌گیری

در این مقاله، یک چارچوب تخصیص منابع مبتنی بر یادگیری تقویتی عمیق برای شبکه‌های یکپارچه زمینی–غیرزمینی (TN-NTN) با کمک RIS فعال و استفاده از CoMP-NOMA ارائه شد. سیستم پیشنهادی RIS فعال که با CoMP-NOMA یکپارچه شده است، در مقایسه با RIS غیرفعال سنتی و پیکربندی‌های مبتنی برOMA ، نرخ مجموع و بهره‌وری انرژی به‌مراتب بالاتری را محقق می‌کند. نتایج شبیه‌سازی نشان می‌دهند که ترکیب RIS فعال با CoMP-NOMA و یادگیری تقویتی عمیق، یک گزینه بسیار کارآمد برای تخصیص منابع و بهبود عملکرد کلی سیستم محسوب می‌شود. این بهبودها به‌طور مشخص در معیارهایی نظیر نرخ مجموع، بهره‌وری انرژی، احتمال قطعی و عدالت میان کاربران قابل مشاهده هستند. در نهایت، این نتایج نشان می‌دهند که استفاده از RIS فعال در کنار CoMP-NOMA و الگوریتم‌های DRL می‌تواند راهکاری مؤثر برای ارتقای عملکرد شبکه‌های نسل آینده باشد. به‌عنوان مسیرهای پژوهشی آینده، می‌توان ادغام فناوری‌های برداشت انرژی با RIS فعال، یا استفاده از ارتباطات موج میلی‌متری (mmWave) و تراهرتز را برای گشودن افق‌های جدید تحقیقاتی مورد بررسی قرار داد.

VI. قدردانی

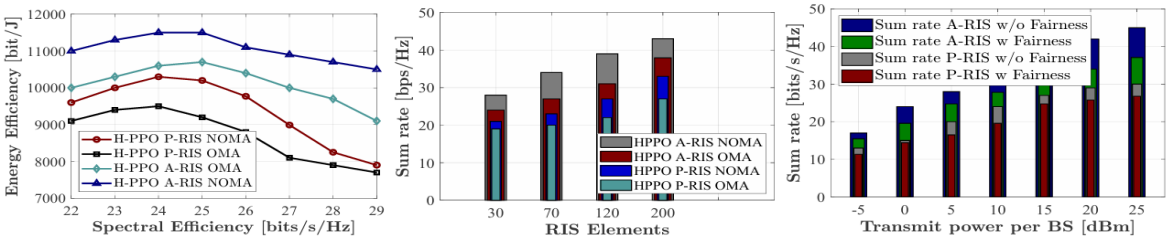
این پژوهش به‌صورت جزئی توسط شورای ملی علوم و فناوری تایوان (National Science and Technology Council — NSTC)، جمهوری چین(.R.O.C) ، تحت گرنت‌های NSTC 113-2622-E-194-011 و NSTC 113-2218-E-194-003 حمایت شده است.

همچنین، این کار تا حدی توسط مؤسسه پیشرفته تولید با نوآوری‌های فناورانه (AIM-HI) در قالب برنامه مراکز پژوهشی حوزه‌های منتخب، در چارچوب پروژه ارتقای آموزش عالی (Higher Education Sprout Project) و با حمایت وزارت آموزش تایوان (MOE) پشتیبانی شده است.



(a) Sum Rate vs transmit power. (b) Reward function vs time steps with normalized RL policies (c) Outage probability vs. transmit power.

Fig. 2: **(a)** Sum rate performance for baseline methods [OMA, Brute Force and RL-NOMA] varying transmit power, **(b)** Training convergence of reward functions for various RL configurations over timesteps, **(c)** Outage probability comparison against transmit power per base station for NOMA and OMA base.



(a) Energy efficiency vs. spectral efficiency. (b) Sum rate vs. RIS elements (c) Sum rate vs. transmit power

Fig. 3: **(a)** Energy efficiency comparison against spectral efficiency between different RIS configuration , **(b)** Sum rate comparison with RIS elements n belong to [30,70,120,200] for different RIS configuration **(c)** Sum rate comparison against baseline methods [DRL NOMA (w and w/o fairness), OMA].

inGLOBECOM2022- 2022 IEEEGlobal Communications Conference,2022,pp.1618–1624.

[8] S. Ammar, C. P. Lau, andB. Shihada, "An in-depth surveyonvir tualization technologies in6g integrated terrestrial andnon-terrestrial networks,"IEEEOpenJournalof theCommunicationsSociety,2024.

[9] M.Bilal,S.F.Zahra,H.Rizwan,T.Umar,S.A.Hassan,H. Jung,and K.Dev, "Singleversus double IRS-assistednetworks:Acomparative analysisusingpracticalphaseshifting,"in2024IEEEWirelessCommu

nicationsandNetworkingConference(WCNC),2024,pp.1–5.

[10] M. Elhattab,M.A.Arfaoui, C.Assi, andA.Ghrayeb, "RIS-assisted joint transmissioninatwo-celldownlinkNOMAcellularsystem,"IEEE JournalonSelectedAreasinCommunications,vol.40,no.4,pp.1270 1286,2022.

[11] M.Obeed,H.Dahrouj,A.M.Salhab,S.A.Zummo,andM.-S.Alouini, "User pairing, link selection, and power allocation for cooperative NOMAhybridVLC/RFsystems,"IEEETransactionsonWirelessCom munications,vol.20,no.3,pp.1785–1800,2021.

[12] Z. Shi, H. Lu, X. Xie, H. Yang, C. Huang, J. Cai, and Z. Ding, "ActiveRIS-aidedEH-NOMAnetworks:Adeepreinforcement learning approach,"IEEETransactionsonCommunications,vol.71,no.10,pp. 5846–5861,2023.

[13] V.Mnih, A. P. Badia,M.Mirza, A. Graves, T. Lillicrap, T.Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning,"inProceedingsofThe33rdInternationalCon ferenceonMachineLearning,M.F.BalcanandK.Q.Weinberger,Eds., vol.48,2016,pp.1928–1937

مراجع

[1] M. Harounabadi and T. Heyn, "Toward integration of 6G-NTN to terrestrialmobilenetworks:Researchandstandardizationaspects,"IEEE WirelessCommunications,vol.30,no.6,pp.20–26,2023.

[2] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurableintelligentsurface-aidedmulti-usernetworks: Interplay betweenNOMAandRIS," IEEEWirelessCommunications, vol. 29, no.2,pp.169–176,2022.

[3] M. Umer,M. A.Mohsin, S. A. Hassan, H. Jung, andH. Pervaiz, "PerformanceanalysisofSTAR-RISenhancedCoMP-NOMAmulti-cell networks,"in2023IEEEGlobecomWorkshops(GCWkshps),2023,pp. 2000–2005.

[4] Y. Ge and J. Fan, "Active reconfigurable intelligent surface assisted secure and robust cooperative beamforming for cognitive satellite terrestrial networks," IEEE Transactions on Vehicular Technology, vol.72,no.3,pp.4108–4113,2022.

[5] R.H.Ratul,M. Iqbal,T.Ashraf,J.-Y.Pan,Y.-H.Wang,andS.-Y.Lien, "Adaptive three layer hybrid reconfigurable intelligent surface for 6g wireless communication: Trade-offs andperformance," in2023 IEEE AsiaPacificConferenceonWirelessandMobile(APWiMob),2023,pp. 232–236.

[6] H.Mei,K.Yang,Q. Liu, andK.Wang, "3D-Trajectory andPhase ShiftDesignforRIS-AssistedUAVSystemsUsingDeepReinforcement Learning,"IEEETransactionsonVehicularTechnology,vol.71,no.3, pp.3020–3029,2022

[7] Z.Zhang,L.Dai,X.Chen,C.Liu,F.Yang,R.Schober,andH.V.Poor, "ActiveRISs:Signalmodeling, asymptoticanalysis, andbeamforming design,"