# GOPT: 트랜스포머 기반 심층 강화 학습을 통한 일반화 가능한 온라인 3D 빈 패킹

헝 시옹, 창롱 궈, 지안 펭, 카이 딩, 웬지에 첸, 쉬총 치우, 롱 바이,
그리고 Jianfeng Xu

***요약-로봇*** **물체 포장은 물류 및 자동화 산업에서 광범위한 실용적 응용 분야를 가지고 있으며, 연구자들은 종종 온라인 3D 빈 포장 문제(3D-BPP)로 공식화합니다. 그러나 기존의 DRL 기반 방법은 주로 제한된 포장 환경에서의 성능 향상에 초점을 맞추고 있으며, 다양한 빈 크기로 특징지어지는 여러 환경에 일반화할 수 있는 능력은 소홀히 하고 있습니다. 이를 위해 저희는 트랜스포머 기반 심층 강화 학습(DRL)을 통해 일반화 가능한 온라인 3D 빈 패킹 접근 방식인 GOPT를 제안합니다. 먼저, 배치 후보와 빈의 표현으로 유한한 하위 공간을 생성하는 배치 생성기(Place- ment Generator) 모듈을 설계합니다. 둘째, 패킹할 항목과 빈 내의 사용 가능한 하위 공간 간의 공간적 상관관계를 파악하기 위해 항목과 빈의 특징을 융합하는 패킹 트랜스포머를 제안합니다. 이 두 가지 구성 요소를 결합하면 다양한 크기의 빈에 대해 추론을 수행할 수 있는 GOPT의 기능이 가능해집니다. 우리는 광범위한 실험을 통해 GOPT가 기준선 대비 우수한 성능을 달성할 뿐만 아니라 일반화 능력도 뛰어나다는 것을 입증했습니다. 또한 로봇을 이용한 배포를 통해 실제 환경에서 이 방법의 실제 적용 가능성을 보여줍니다.**

***색인*** **용어-조작 계획, 강화 학습, 로봇 포장.**

## I. 소개

세계 무역 및 전자상거래 시장의 번영과 함께 창고 자동화는 빠르게 발전해 왔습니다.
에 대한 관심이 높아지고 있습니다. 창고 내 효율적인 물건 배치

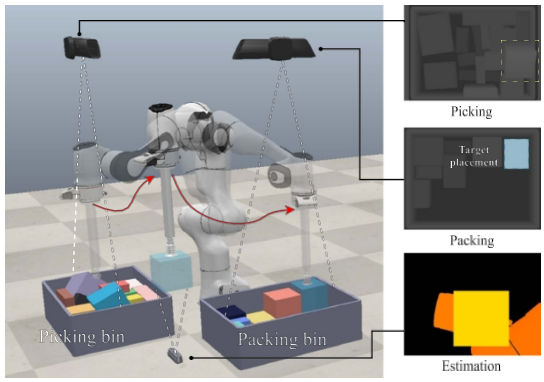



그림 1. 로봇 피킹 및 포장 파이프라인. 왼쪽: 로봇이 어지럽게 쌓인 상자에서 물건을 무작위로 골라 컴팩트하게 포장하고 있으며, 세 대의 RGB-D 카메라가 장착되어 있습니다. 오른쪽: 두 대의 오버헤드 카메라가 각각 두 개의 쓰레기통의 상태를 관찰하고, 한 대의 위쪽 카메라가 피킹된 물품의 치수를 추정합니다.

최적의 포장 전략을 통해 노동력 감소, 비용 절감 등 다양한 이점을 얻을 수 있습니다[1].

그림 1은 로봇 팔을 이용한 물품 피킹 및 포장의 예시를 보여줍니다. 여기서는 로봇 피킹이 잘 구현되어 있다고 가정합니다. 연구자들은 일반적으로 로봇 포장의 배치 문제를 온라인 3D 빈 포장 문제(3D-BPP)로 공식화하여 해결해 왔습니다[2], [3]. 고전적인 조합 최적화 문제 중 하나인 3D-BPP는 공간 활용을 극대화하기 위해 알려진 정육면체 아이템 세트를 축 정렬 방식으로 빈에 배치하는 문제입니다. 그러나 모든 항목을 관찰하고 이에 대한 완전한 지식을 얻는 것은 많은 실제 시나리오에서 어려운 일입니다. 온라인 3D-BPP는 들어오는 품목만 관찰하면서 품목을 하나씩 포장하는 3D-BPP의 보다 실용적인 변형입니다.

지식이 제한되어 있기 때문에 온라인 3D-BPP는 정확한 알고리즘으로 해결할 수 없습니다[4]. 연구자들은 이전에 인간 포장업자의 경험을 추상화하여 설계된 문제에 대한 욕심 많은 목표를 가진 휴리스틱을 개발하는 데 주력해 왔습니다[5]. 그러나 이러한 휴리스틱은 직관적이기는 하지만 일반적으로 차선의 솔루션을 산출합니다. 최근 몇 년 동안 심층 강화 학습(DRL)[2], [3], [6], [7]을 통한 온라인 3D-BPP 해결에 대한 연구가 활발히 진행되고 있으며, 실제로 DRL 기반 방법은 인상적인 성능을 보여주고 있습니다. 그럼에도 불구하고 훈련 과정에서 종종 융합에 도달하는 데 어려움을 겪고 있으며[2], [8], 이러한 방법은 여러 분야에 걸쳐 효과적으로 일반화하기 어렵다는 점에 주목할 필요

2024년 6월 22일 접수, 2024년 9월 7일 승인. 발행일 2024년 9월 25일, 현재 버전 발행일 2024년 10월 10일. 이 글은 검토자의 의견을 평가한 후 부편집장 Heping Chen과 편집장 Chao-Bo Yan이 게재를 다시 추천했습니다. 이 연구는 중국 국가 중점 R&D 프로그램의 지 원 으 로 2022YFB4700300 보조금을 받아 수행되었습니다. *(교신저자: 지안펑 쉬).*

헝 시옹, 창롱 궈, 지안 펭, 롱 바이 씨는 중국 우한 430074의 화중과학기술대학교 기계공학부 지능형 제조 장비 및 기술 국가 중점 연구소에 소속되어 있습니다(이메일: xiongheng@hust.edu.cn; guochangrong@hust.edu.cn; peng_jian@hust.edu.cn; bailong@hust.edu.cn).

카이 딩과 쉬총 치우는 중국 상하이 200335에 있는 보쉬 코퍼레이트 리서치(이메일: kai.ding@cn.bosch.com; xuchong.qiu@cn.bosch.com)에 있습니다. 중국 포산 528300의 Midea 그룹 하이엔드 헤비로드 로봇 국가 핵심 연구소에 소속되어 있습니다.

중국 포산 528311, 연구 센터(이메일: chenwj42@midea.com). 지안펑 쉬는 지능형 제조 국가 핵심 연구소에 소속되어 있습니다.

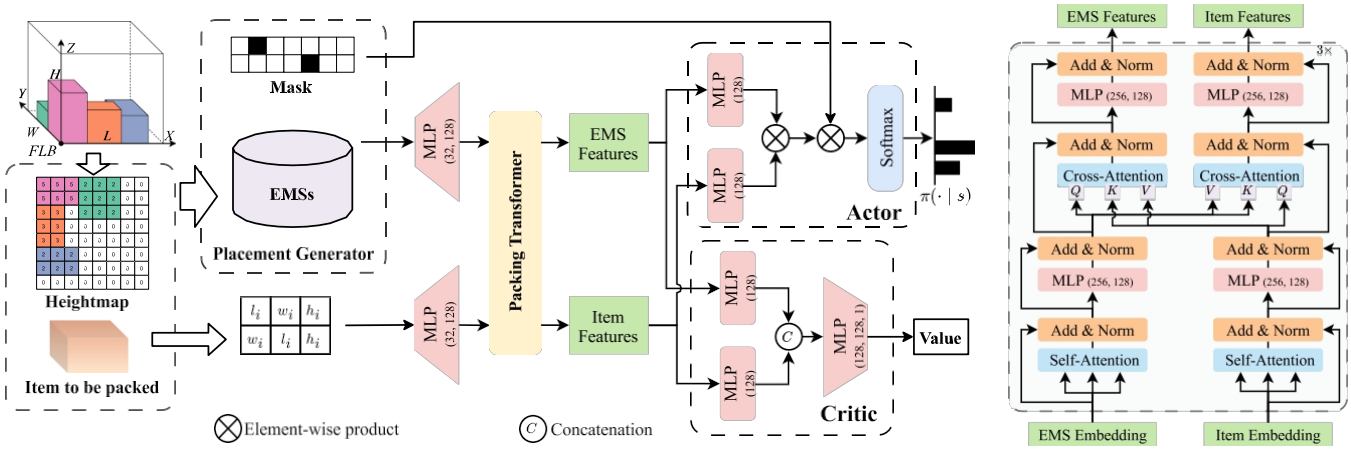
중국 우한 430074에 위치한 화중과학기술대학교 기계과학공학부 장비 및 기술, 그리고 중국 우시 214174에 위치한 HUST-Wuxi 연구소와 함께합니다(이메일: jfxu@hust.edu.cn).

소스 코드는 https://github.com/Xiong5Heng/ GOPT에서 공개됩니다.

이 서한에는 저자가 제공한 추가 다운로드 자료가
https://doi.org/10.1109/LRA.2024.3468161 에서 제공됩니다.

디지털 객체 식별자 10.1109/LRA.2024.3468161

가 있습니다.




(a) The framework of GOPT (b) Packing Transformer

그림 2. 방법 개요. (a) GOPT에서 입력은 포장할 품목과 빈의 현재 높이맵으로 구성되며, 각 셀의 값은 각 높이를 나타냅니다. 배치 생성기를 사용하여 각 EMS와 항목의 선택적 방향 사이의 쌍별 액션 마스크와 함께 EMS 세트가 생성됩니다. 그 후 EMS와 아이템을 별도로 인코딩한 다음 패킹 트랜스포머를 사용하여 특징을 융 합 하 고 , 이 중 출력을 액터 및 비평 네트워크에 공급하여 모든 액션의 로그를 생성하고 상태값 함수를 추정합니다. (b) 제안된 패킹 트랜스포머의 세부 사항을 나타냅니다. 이 트랜스포머는 3개의 스택형 블록으로 구성되며, 각 블록에는 2개의 자체 주의 레이어와 2개의 교차 주의 레이어가 포함되어 있습니다.

다양한 포장 시나리오, 특히 다양한 빈 치수가 특징인 포장 시나리오를 고려해야 합니다. 이러한 한계는 일반적인 사용 사례에서 DRL의 광범위한 적용 가능성을 상당히 축소시킵니다. 보다 구체적으로, 현재의 최신 DRL 기반 방법은 학습된 것과 동일한 크기의 빈에 대해서만 추론을 수행할 수 있습니다 [3], [9]. 학습된 모델은 크기가 다른 빈으로 이전할 수 없습니다. 또한 이러한 방법에서 패킹 작업 공간 크기가 빈 크기에 내재적으로 의존하기 때문에 특히 더 큰 빈을 다룰 때 모델 수렴에 상당한 어려움이 있습니다[10].

앞서 언급한 한계에 착안하여 그림 2와 같이 트랜스포머 기반 DRL을 통해 일반화 가능한 온라인 3D 빈 패킹 접근 방식인 GOPT를 제안합니다. GOPT에서 *배치 생성기(PG)* 모듈은 먼저 휴리스틱을 채택하여 임대 빈 내에서 고정 길이의 자유 하위 공간 집합을 배치 후보로 생성하여 패킹 작업 공간의 크기를 제어할 수 있도록 합니다. 배치 후보와 포장할 물품은 모두 집합적으로 마르코프 결정 과정(MDP)의 상태로 정의됩니다. 그런 다음 GOPT는 *패킹 트랜스포머* 모듈을 통합하는 새로운 패킹 정책 네트워크를 통합합니다. 이 모듈은 현재 항목과 사용 가능한 하위 공간 사이의 공간적 상관관계와 PG 모듈에서 파생된 이러한 하위 공간 간의 관계를 본질적으로 식별하여 GOPT의 일반화 가능성을 향상시킵니다. 패킹 트랜스포머는 자기 주의 레이어와 양방향 교차 주의 레이어를 사용하여 강화 학습 정책의 입력으로 특징을 추출합니다.

실험 결과, 공간 활용률과 포장된 물건의 수 측면에서 최신 포장 방법보다 성능이 뛰어난 것으로 나타났습니다. 저희가 아는 한, 고성능을 유지하면서 훈련된 모델을 통해 다양한 쓰레기통에서 추론할 수 있는 일반화 기능을 제공하는 것은 저희의 연구가 처음입니다. 또한 포장 계획 방법을 로봇 매니퓰레이터에 배포하여 실제 환경에서 실제 적용 가능성을 입증했습니다.

요약하자면, 저희의 주요 공헌은 다음과 같습니다: (1) 패킹 성능과 일반화를 확대하는 온라인 3D-BPP의 새로운 방법인 GOPT, (2) 패킹 작업 공간을 변조하고 빈의 상태를 나타내는 배치 생성기 모듈, (3) 현재 항목과 사용 가능한 하위 공간 간의 관계 및 하위 공간 간의 상호 관계를 캡처하는 패킹 트랜스포머라는 네트워크입니다;
(4) GOPT와 기준선을 비교하는 광범위한 실험 평가.

## II. 관련 작업

3D-BPP는 고전적인 최적화 문제이며 NP가 매우 어려운 것으로 알려져 있습니다[11]. 여기서는 관련 휴리스틱 및 DRL 기반 방법을 간략하게 살펴봅니다.

### *A. 휴리스틱 방법*

초기 연구는 주로 단순성을 위해 효율적인 휴리스틱을 설계하는 데 중점을 두었습니다. 연구자들은 퍼스트 핏[12], 베스트 핏[13], 가장 깊은 하단-왼쪽 채우기[14] 등 작업자의 경험에서 추출한 몇 가지 패킹 규칙을 정의하려고 시도했습니다. 코너 포인트(CP)[15], 극한 포인트(EP)[16], 빈 최대 공간(EMS)[17], 내부 모서리 포인트(ICP)[18]는 휴리스틱 방법을 개선하기 위해 물품을 포장할 수 있는 잠재적 여유 공간을 나타내기 위해 노력합니다. 예를 들어, Ha 등[5]은 포장할 물품의 면과 EMS의 면 사이의 여백을 최소화하기 위해 하나의 EMS를 선택하는 OnlineBPH를 제안합니다. 야림캠 등[19]은 Irace 매개변수 튜닝 알고리즘을 사용하여 정책 행렬로 표현된 휴리스틱을 제공합니다[20]. Wang 등[21]은 점유량을 최소화하는 배치를 선호하는 높이맵 최소화(HM)를 제안합니다. Shuai 등은 현실 세계에서 발생하는 불확실성을 완화하기 위해 변형된 상자를 가깝게 쌓아 안정성을 높입니다[22]. Hu 등은

잠재적으로 큰 미래의 아이템을 포장하기 위해 사용 가능한 빈 공간을 최적화하는 MACS(Maximize-Accessible-Convex-Space) 전략[23]. 이러한 방법은 직관적이고 효과적이지만, 수작업으로 만든 규칙에 의존하고 다양한 문제 설정에서 일관되게 우수한 성능을 발휘하기에는 역량이 부족합니다. 우리의 작업은 휴리스틱의 빈 공간 표현을 활용하지만, DRL을 사용하여 도메인 전문 지식의 제한을 받지 않고 패킹 패턴을 학습합니다.

### B. *DRL 기반 방법*

DRL은 특정 조합 최적화 문제를 해결할 수 있는 가능성을 보여주었습니다[24], [25]. 따라서 최근 3D-BPP를 해결하기 위해 DRL을 사용하는 추세가 있습니다. Que 등[26]은 위치, 항목 선택, 방향의 하위 작업을 순차적으로 처리하기 위해 트랜스포머 구조의 DRL을 사용하여 높이가 가변적인 오프라인 3D-BPP를 해결합니다. 대신 온라인 3D-BPP에 집중하여 위치와 방향을 동시에 결정합니다. 우리가 아는 한, Deep-Pack[27]은 2D 온라인 포장 문제를 해결하기 위해 DRL 기반 모델을 사용한 최초의 사례이며, 온라인 3D-BPP로 확장할 가능성이 있습니다. 빈의 현재 상태를 보여주는 이미지를 입력으로 받아들어오는 물품을 포장하기 위한 픽셀 위치를 출력합니다. Verma 등[6]은 검색 휴리스틱과 DRL을 결합하여 빈의 수와 크기에 상관없이 문제를 해결하기 위한 2단계 전략을 제안합니다. Zhao 등[2], [10]은 문제를 제약된 MDP로 공식화하고 ACKTR 방법[28]을 채택하여 CNN 기반 DRL 에이전트를 훈련합니다. 2]에서 DRL 에이전트는 행위자, 비평가, 예측자로 구성되어 각각 행동 확률, 값, 타당성 마스크를 추정합니다. 그런 다음 패킹 액션을 길이와 너비 차원과 방향으로 분해하여 액션 공간을 줄이는 방식으로 개선합니다[10]. 이후 휴리스틱 검색 규칙에 기반한 패킹 구성 트리(PCT)를 도입하고 이를 DRL 에이전트에 통합합니다[8]. 이 에이전트는 그래프 주의 네트워크[29]를 정책으로 사용하며 ACKTR로 훈련됩니다. 휴리스틱과 DRL의 시너지 효과를 조사하기 위해 Yang 등[7]은 휴리스틱 보상을 활용하여 DRL 에이전트가 더 나은 성능을 발휘하도록 지원하는 PackerBot을 제안합니다. Xiong 등[3]은 후보 맵 메커니즘을 도입하여 탐색의 복잡성을 줄이고 A2C로 훈련된 CNN 기반 DRL 에이전트의 성능을 개선합니다[30]. 이러한 방법은 일반적으로 항목과 빈의 특징을 직접 연결하여 정책을 학습합니다. 이와는 대조적으로 GOPT는 먼저 빈 내의 자유 하위 공간을 제안하고 수정된 트랜스포머를 사용하여 이러한 공간 간의 관계와 현재 항목과의 관계를 파악합니다. 이 방법은 다양한 포장 환경에서 일반화 가능성을 보장합니다.

## III. 방법론

### A. *문제 설명*

그림 1과 같이 로봇은 다양한 치수의 상자 모양의 물건이 모여 있는 비정형 더미에서 무작위로 물건을 골라냅니다. 모든 항목에 대한 완전한 지식은 미리 알 수 없습니다. 카메라 한 대가 피킹된 물품의 치수를 측정한 다음 포장 상자에 넣습니다. 이
특정 시나리오는 온라인 3D-BPP로 특징 지을 수 있습니다. 온라인 3D-BPP는 의 목표는 가능한 한 많은 항목을 보관함에 넣고 보관함의 공간 활용도를 극대화하는 것입니다.

그림 2(a)와 같이 차원 ($L, W, H$)을 가진 구간차원의 앞-왼쪽-아래(FLB) 정점을 원점(0, 0, 0)으로, 길이, 너비, 높이 방향을 각각 $X$, $Y$, $Z$ 방향으로 정의합니다. 항목의 경우 ($X_T$, $Y_T$, $Z_T$)는 치수가 ($L_T$, $W_{(T)}$) , $H_T$인 t *번째*항목의 FLB 좌표를 나타냅니다.

로봇 포장 작업에서는 다음과 같은 물리적 제약을 고려해야 합니다.

*직각 배치:* 항목이 휴지통에 직각으로 배치되고 측면이 휴지통의 측면과 정렬됩니다.

*선택적 방향*: 항목은 똑바로 세워서 배치되며, 첫 번째 제약 조건과 함께 항목의 평면 내 수직 방향은 0° 또는 90° 중 두 가지로만 지정할 수 있습니다.

*정적 안정성*: 포장 과정에서 물품은 중력 및 물품 간 힘에 의해 안정적으로 유지되어야 합니다. 계산 효율성을 위해 기하학적 중심을 바닥에 투영했을 때 해당 품목의 모든 수평 지지점의 볼록한 선체로 형성된 지지 다각형 안에 해당 품목이 속하면 안정된 것으로 간주합니다[23].

### B. *배치 생성기*

포장할 선택한 물품의 수평 위치($x_t$, $y_t$)와 해당 물품이 빈에 배치될 방향을 예측합니다. 수직 위치 $z_t$는 중력으로 인해 가장 낮은 배치 위치에 의해 분석적으로 결정됩니다. 앞서 언급했듯이 하나의 항목에 대해 두 가지 방향이 가능합니다. 따라서 치수가 ($L$, $W$, $H$) 인 빈에 품목을 배치할 때 총 배치 가능 개수는 $L \times W \times 2$ 개가 됩니다 [2]. 한편으로, 이 수량은 빈 크기가 커질수록 기하급수적으로 증가하기 때문에 순차 결정 특성을 가진 포장 문제에서는 견딜 수 없습니다. 다른 한편으로, 이 배치 세트 내에서 포장할 품목 중 일부는 불가피하게 비생산적일 수밖에 없습니다.

잠재적으로 큰 배치 검색 공간을 제한하기 위해, 들어오는 항목과 현재 빈 구성을 기반으로 유한하고 효율적인 배치 하위 집합을 생성하는 배치 생성기(PG) 모듈을 설계합니다. 먼저 하이트맵을 활용하여 보관함의 실시간 상태를 명시적으로 표현합니다. 이전 항목에 대한 계획된 배치를 표현으로 활용하는 다른 방법[8]은 피드백 및 폐루프 제어가 부족합니다. 반면 하이트맵은 실제 로봇 포장 작업에서 PG를 배치할 때 카메라로 캡처한 시각적 관찰을 통해 편리하게 도출할 수 있습니다. 빈 공간의 빈 공간을 관리하기 위한 빈 최대 공간(EMS) 방식[17], [31]에서 아이디어를 얻어 현재 상태를 기반으로 후보 배치를 계산합니다. 구체적으로 하이트맵의 $X$ 및 $Y$ 방향을 따라 높이 변화를 감지하여 코너 포인트를 식별합니다. 그런 다음 각 코너에서 단위 직사각형을 확장하고 더 높은 고도를 만나면 멈추는 방식으로 EMS를 생성합니다(그림 3). 각 EMS는 그림 3(c)에 표시된 것처럼 FLB 꼭지점과 그에 대응하는 반대쪽 꼭지점으로 정의할 수 있습니다. 결과 6차원 벡터는 구간차원의 차원에 관계없이 [0, 1]로 정규화됩니다. 제어 가능한 크기의 EMS 하위 집합을 얻고 높이 값에 따라 순위를 매깁니다,

i=1

$E_{(i)\ f}{}^{N}$으로 표시됩니다. . 마지막으로 포장할 품목이 주어지면

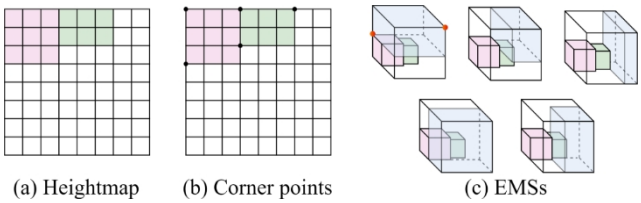



그림 3. EMS 생성 절차 그림. (a) 두 개의 아이템이 배치된 예시 장면에서 하이트맵은 각 그리드 셀에 쌓인 아이템의 현재 높이를 나타냅니다. (b) 이 하이트맵에서 5개의 모서리 점(검은색 점)이 감지됩니다. (c) 이 점을 기준으로 빈 내에서 해당하는 가장 큰 사각형(파란색)이 생성되며, 이것이 바로 EMS입니다. 첫 번째 EMS를 예로 들면, 파란색 직사각형의 빨간색 꼭지점 두 개로 정의됩니다.

섹션 III-A에 따라 각 EMS의 타당성을 확인하고 각 EMS와 방향 사이에 쌍으로 마스크를 생성합니다. 빈에 상품을 포장할 때 적절한 EMS와 방향을 선택하고 상품과 EMS의 FLB 정점을 정렬합니다.

### C. 강화 학습 공식

DRL 문제는 일반적으로 마르코프 결정 과정(MDP)으로 모델링됩니다. 여기서 매개변수 $\times S, A, P, R, \gamma)$가 있는 MDP를 사용하여 패킹 환경을 특성화하는데, 여기서 $S$는 상태 공간, $A$는 행동 공간, $P : S \times A \times S \to [0,+\infty)$는 전이 확률, $R : S \times A \to$ R은 스칼라 보상 함수, $\gamma \in (0, 1]$은 DRL에서 단기 및 장기 보상 간의 균형을 맞추는 할인 계수를 나타냅니다. 강화 학습 알고리즘은 상태 $s_t$가 주어졌을 때 행동을 선택할 확률을 결정하는 정책 $\pi : S \times A \to$ R을 학습하는 것을 목표로 합니다. 정책의 목표는 한 에피소드에 대한 누적 할인 보상을 극대화하는 것으로, $\sum_{(t)} \gamma^{(t)} r_t$로 표현되며 여기서 $t$는 시간 단계를 나타내고 $r_t$, $a_{(t)}$, $s_t$는 각각 시간 단계 $t$에서의 보상, 행동, 상태를 나타냅니다. 다음에서는 온라인 3D-BPP를 DRL 학습을 위한 MDP로 공식화합니다.

*상태:* 상태: 각 시간 단계 $t$에서 정책은 포장할 수신 항목 $s_{t,item}$과 현재 빈 구성 $s_{t,bin}$을 포함하는 상태 $s_{(t)}$를 수신합니다. 첫 번째 부분에서는 항목의 크기($l_t$, $w_t$, $h_t$)가 필수적입니다. 일부 연구[3], [7]에서는 이 3차원 벡터를 항목 표현으로 명시적으로 사용하는 반면, 다른 연구에서는 신경망 설계의 편의성을 위해 3채널 맵을 선호합니다[2], [9]. 지도 표현에서 각 채널은
에 각각 $L_T$, $W_T$, $H_T$가 할당됩니다. 지오메트리와 선택적 방향을 모두 고려하기 위해 항목 대표

2× 3 행렬인 resentation, $s_{(t,item)}$ = $\begin{matrix} w_t & h_{(t)} \\ w_t & l_t \end{matrix}$

여기서 ($L_T$, $W_T$, $H_T$)와 ($W_T$, $L_T$, $H_T$)는 항목을 0°와 90° 회전한 후의 치수를 나타냅니다. 두 번째 부분의 경우 기존 방법에는 하이트맵 [3], 패킹된 항목 목록 [8], 가중치가 적용된 3D 복셀 그리드 [9] 등이 있습니다. 우리는 제안된 PG(섹션 III-B)를 활용하여 빈의 구성으로 배치 제약 조건을 만족하는 EMS 시퀀스를 생성하기로 결정했습니다. 이 시퀀스는 더미 EMS를 사용하여 고정 길이 $N$으로 패딩 또는 클리핑됩니다(즉, $s_{t,bin} = \{E_{(i)}\}_{i=1}^{N}$.

*액션:* 패킹 상태 $s_t = (s_{t,item}, s_{t,bin})$가 주어지면, 액션 $a_{(t)}$는 사용 가능한 EMS 시퀀스에서 현재 항목의 방향과 EMS를 모두 선택하는 작업을 포함합니다. 액션 공간 $A$의 크기는 빈 차원에 관계없이 시퀀스의 길이와 선택적 방향의 수(예: $|A| = 2N$)에만 의존합니다. 훈련 중에는 액션 $\pi(\cdot | s_t)$에 대한 확률 분포에 따라 액션 $a_{(t)}$를 선택하며, 여기서 $\cdot$는 $s_t$에서 가능한 모든 배치의 집합을 나타냅니다. 테스트하는 동안 $\pi(\cdot | s_t)$에서 최대 확률을 가진 배치를 선택하여 결정론적 방식으로 액션을 선택합니다. EMS와 방향 사이에 쌍별 액션 마스크를 적용하는 확률 분포는 제약 조건을 충족하는 EMS가 없는 한 정책에서 유효한 액션을 샘플링하도록 보장한다는 점에 유의하세요.

*상태 전환:* 이 설정에서 전환 모델은 결정론적인 것으로 합산되며, 이는 특정 쌍($s_t$, $a_t$)이 일관되게 동일한 후속 상태 $s_{(t)(+1)}$로 이어진다는 것을 의미합니다.

*보상:* 포장 문제의 목표는 빈의 공간 비율을 최대화하는 것입니다. 따라서 보상은 공간 활용률의 단계적 향상으로 공식화하며, 이는 $r_{(t)}$로 표시됩니다. $= \frac{(L)_{(T)(-)} (W)_{(T)(-)} (H)_{(T)}}{L \cdot W \cdot H}$. 이 밀도 높은 보상은 DRL 에이전트가 다음을 수행하도록 장려합니다.
에피소드에 더 많은 단계를 추가하여 더 많은 아이템을 포장하고 공간 활용도를 높일 수 있습니다.

### D. 네트워크 아키텍처

선택한 아키텍처가 다양한 환경에서 에이전트의 학습 및 일반화 기능에 영향을 미치기 때문에 DRL 에이전트를 위한 신경망 아키텍처의 설계는 중요합니다. 간단한 네트워크는 빈과 항목 표현[2] 또는 임베딩[7]을 연결하는 것입니다. 그러나 이 방법은 컨볼루션 및 선형 레이어 크기가 빈의 크기에 따라 달라지는 모델을 생성하므로 학습된 모델을 여러 빈에 적용하기에는 실용적이지 않습니다.

일반화의 문제를 극복하기 위해, 저희는 항목과 빈의 부분 공간 사이의 상관관계에 초점을 맞춘 주의 기반 네트워크 아키텍처를 제안합니다. 그림 2(a)에서 볼 수 있듯이, 이 아키텍처는 패킹 트랜스포머, 액터 네트워크, 비평가 네트워크의 세 가지 주요 구성 요소로 이루어져 있습니다. 우리의 네트워크는 빈 표현 $s_{t,bin} \in R^{N\times 6}$(즉, PG의 EMS 시퀀스)와 항목 표현 $s_{t,item} \in R^{2\times 3}$(즉, 항목의 차원)을 입력으로 받습니다. 그런 다음 이러한 입력은 LeakyReLU 활성화 기능을 갖춘 2계층 선형 네트워크인 다층 퍼셉트론(MLP)에 의해 개별적으로 처리됩니다. 임베딩 차원
섹션과 항목 모두 128로 설정됩니다. 이후
그런 다음 설계된 임베딩을 사용하여 임베딩에서 피처를 추출합니다.

크로스 모달리티 학습에서 영감을 받은 패킹 트랜스포머(Packing Transformer)는 언어와 시각 [32]. 그런 다음 EMS와 아이템의 특징을 액터 네트워크에 입력하여 잠재적 행동의 확률 분포를 생성하고 비평가 네트워크에 입력하여 현재 상태를 기반으로 예상 누적 보상을 추정합니다.

**패킹 트랜스포머는** 그림 2(b)에 자세히 설명되어 있습니다. 이는 각각 2개의 셀프 어텐션 레이어, 1개의 양방향 크로스 어텐션 레이어, 2개의 MLP 블록으로 구성된 4개의 동일한 인코더 블록을 여러 개(실제로는 3개) 쌓아서 구성됩니다.

128, 128} 개의 뉴런으로 구성된 레이어입니다. 양방향 교차 주의 레이어는 두 개의 단방향 교차 주의 레이어로 구성되며, 하나는 EMS에서 항목으로, 다른 하나는 항목에서 EMS로 구성됩니다. 잔여 연결과 레이어 정규화(Norm)는 각 레이어 뒤에 적용됩니다. 자체 주의 레이어는 EMS 또는 항목 차원 간의 내재적 연결을 설정하는 데 중요한 역할을 하는 반면, 양방향 교차 주의 레이어는 서로 간의 내부 관계를 쉽게 발견할 수 있게 해줍니다.

**액터 네트워크와 비평가 네트워크는** 모두 그림 2(a)에 표시된 MLP 레이어로 구현됩니다. 액터 네트워크에서는 EMS와 항목 특징이 모두 MLP를 통해 처리되고 그 결과를 곱하여 액션의 점수 맵을 계산합니다. 그 다음에는 실행 불가능한 액션을 제거하기 위해 액션 마스크와 요소별 곱셈을 수행합니다.

### E. 교육 방법

우리는 제안된 GOPT를 훈련하기 위해 근거리 정책 최적화(PPO) 알고리즘 [33]을 사용합니다. PPO는 환경과의 상호작용을 통해 데이터를 수집하고 각 반복에서 대략적으로 최대화되는 다음 목표를 최적화하는 번갈아 가며 사용하는 널리 사용되는 온-정책 강화 학습 알고리즘입니다:

$$L(\vartheta)= \hat{\mathrm{E}}_{(t)}\,[L^{(C)\,(L)\,(I)\,(P)\,(\theta)}-c_{(1)}\,L^{(VF)\,(\theta)}+c_{(2)}\,S(\pi_{(\theta)\,(}\cdot\mid s_{t)})] \quad (1)$$

여기서 θ는 네트워크 파라미터, $c_1$, $c_2$는계수, $L^{CLIP)\,(\theta}$는 클리핑된 대리 대물렌즈, $L^{(VF)\,(\theta)는}$클립된 대물렌즈를 나타냅니다.

는 가치 함수의 제곱오차 손실이고, $S$는 정책의 엔트로피를 나타냅니다. 구체적으로 대리 목표는 다음과 같이 정의됩니다:

$$L^{CLIP}= \hat{\mathrm{E}}\,[\min(p_t(\vartheta)\hat{A}_t,\ \mathrm{clip}(p_t(\vartheta),\ 1-\epsilon,\ 1+\epsilon)\hat{A}_t)] \quad (2)$$

여기서 $p_t(\vartheta)=\frac{\pi_{(\theta)\,(a_t\mid s_t())는}}{\pi_{(\theta)\,(구)\,(a_t\mid s_t)}}$다음 사이의 동작 확률 비율입니다.

는 현재 정책과 이전 정책의 차이, $\hat{A}_{t는}$ 일반화된 이득 추정기(GAE)[34] 방법을 사용하여 계산하는 이득 함수의 추정치, ϵ는 업데이트 양을 제한하고 학습 절차를 안정화하는 데 사용되는 클리핑 비율을 나타냅니다.

## IV. 실험

### A. 구현 세부 정보

우리의 방법은 PyTorch를 활용하여 구현되며 정책 학습을 위해 Tianshou 프레임워크[35] 내의 PPO 알고리즘을 채택합니다. 각 패킹 단계에서 최대 EMS 수는 80개로 설정됩니다. 1000개의 에포크에 대해 정책을 학습하고 모든 에포크에서 128개의 병렬 환경에 걸쳐 총 40,000개의 환경 단계를 수집합니다. 정책 업데이트는 640개의 환경 단계(5× 128단계로 계산)마다 발생하며, 배치 크기는 128입니다. $7\times 10^{-5}$에서 시작하는 선형 하강 학습률 스케줄러와 결합된 Adam 옵티마이저가 최적화에 활용됩니다. PPO 손실 계산의 경우, 가치 및 엔트로피 손실 계수 $c_1$, $c_2$는 각각 0.5, 0.001이고 클립 비율 ϵ는 0.3입니다. 할인 계수 $\gamma$는 미래 보상과 즉각적인 보상을 동등하게 고려하기 위해 1로 설정됩니다. 정책 업데이트의 경우, $\lambda_{GAE}=$ 0.96의 GAE를 사용합니다.

표 I
10× 10× 10 빈에 대한 성능 비교
절제 연구 결과

| Method | *Uti* | *Sta* | *Num* |
|---|---|---|---|
| ***Heuristic*** | | | |
| OnlineBPH [5] | 51.6% | 0.142 | 20.5 |
| Best Fit [16] | 57.9% | 0.124 | 22.9 |
| MACS [23] | 50.6% | 0.171 | 19.6 |
| HM [21] | 56.5% | 0.105 | 22.1 |
| ***DRL-based*** | | | |
| Zhao et al. [2] | 70.9% | 0.079 | 27.5 |
| PCT [8] | 72.7% | 0.073 | 28.1 |
| Xiong et al. [3] | <u>73.8%</u> | **0.068** | <u>28.3</u> |
| GOPT (ours) | **76.1%** | <u>0.070</u> | **29.6** |
| ***Ablation studies*** | | | |
| GOPT w/o PG | 70.6% | 0.086 | 27.5 |
| GOPT w/o IR | 73.2% | 0.078 | 28.5 |
| GOPT w/o PT | 67.1% | 0.085 | 26.2 |
| GOPT-MLP | 67.8% | 0.079 | 26.4 |
| GOPT-GRU | 68.7% | 0.082 | 26.9 |

**Bold** indicates the best and <u>underline</u> indicates the second best for that metric.

. 정책 교육은 엔비디아 지포스 RTX 3090과 인텔 코어 i7-14700K CPU가 탑재된 컴퓨터로 진행되며, 약 6시간 만에 컨버전스를 기초부터 완성할 수 있습니다.

실험적 검증을 위해, 저희는 DRL 에이전트를 훈련하고 평가하기 위해 RS 데이터 세트[2]를 활용합니다. 빈 크기 $L\times W\times$ H는 10× 10× 10으로 설정하고, 크기 $\frac{(\min()\,(L,W,H)\,())}{10}\le L_{(}$, $W_{(T)}$, $H_T\le(\frac{\min()\,(L,W,H)\,())}{10}$. $^2$ 데이터 세트는 125가지 유형의 이질적인 항목으로 구성되어 있으며, 실제 시나리오의 가변성을 반영하기 위해 훈련 중에 부트스트랩 샘플링을 통해 시퀀스를 동적으로 생성합니다. 평가를 위해 추가로 1000개의 인스턴스 세트가 생성됩니다.

$_t$ 로 설정하고 평균 성능을 기록합니다.

### B. 성능 평가

*1) 베이스라인:* 저 희 방법의 우수성을 설명하기 위해 공개적으로 구현된 대표적인 방법을 베이스라인으로 선정했습니다. 이러한 방법을 두 가지 그룹으로 분류합니다. 첫 번째 그룹은 네 가지 휴리스틱 방법으로 구성됩니다: OnlineBPH [5], 가장 낮은 극한점에 있는 항목을 패킹하는 EP [16]에 기반한 베스트 핏, MACS [23], HM [21]. 두 번째 방법은 세 가지 DRL 기반 방법으로 구성됩니다: Zhao 등 [2], PCT [8], Xiong 등 [3]. 모든 방법은 공정하고 엄격한 비교를 보장하기 위해 동일한 데스크톱 컴퓨터에서 구현 및 실행됩니다. 또한, DRL 기반 방법은 훈련 불균형 편향을 제거하기 위해 동일한 수의 단계, 특히 4천만 단계로 처음부터 훈련됩니다.

*2) 결과:* 빈의 평균 공간 활용도(*Uti*), 포장된 물품의 평균 개수(*Num*), 공간 활용도의 표준편차(*Sta*)라는 세 가지 지표를 사용하여 이러한 방법의 포장 성능을 평가했으며, 후자는 모든 경우에 걸쳐 방법의 안정성을 평가합니다. 표 I에 제시된 정량적 비교 결과, 우리 방식이 모든 기준선(*Uti* 및 *Num*)을 능가하는 것으로 나타났습니다. 이 결과는 우리 방식이 우수한 물품 포장과 빈 공간의 효율적인 활용을 달성한다는 것을 보여줍니다. 주목할 만한 점은 *Sta* 측면

에서 두 번째로 높은 성능을 달성했다는 점입니다,

표 II
다양한 크기의 빈에 대한 일반화 성능

| Method | Bin-10 | | Bin-30 | | Bin-50 | | Bin-100 | |
|---|---|---|---|---|---|---|---|---|
| | *Uti* | *Num* | *Uti* | *Num* | *Uti* | *Num* | *Uti* | *Num* |
| Zhao et al. [2] | 70.9% | 27.5 | 72.4% | 27.9 | 51.7% | 20.6 | / | / |
| Zhao et al. [10][1] | 70.1% | 27.1 | 71.7% | 27.7 | 72.6% | 28.1 | 71.3% | 27.6 |
| PCT [8] | 72.7% | 28.1 | 73.1% | 28.1 | 70.1% | 27.2 | 72.7% | 27.9 |
| Xiong et al. [3] | 73.8% | 28.3 | 75.6% | 28.9 | 75.3% | 28.8 | 73.8% | 28.2 |
| GOPT | **76.1%** | **29.6** | **76.0%** | **29.5** | 75.7% | **29.4** | 75.7% | 29.4 |
| GOPT (Bin-10)[2] | **76.1%** | **29.6** | **76.0%** | 29.2 | **75.8%** | 29.2 | **76.3%** | **29.6** |

[1]Results are copied directly from [10] since the code is not available.
[2]GOPT (Bin-10) refers to the GOPT policy trained in Bin-10, which we directly apply to four environments to obtain testing results. In contrast, the other four methods, along with GOPT, require separate training and testing in these environments.

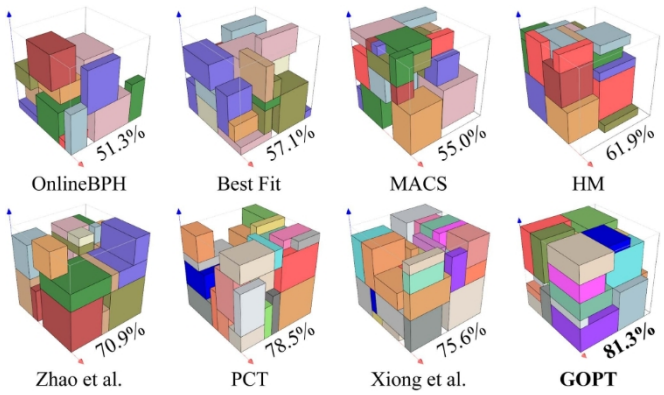



그림 4. 에서 항목 시퀀스에 대한 다양한 방법의 시각화 결과 10× 10× 10 bin. 각 빈 옆의 숫자는 *Uti*의 값을 나타냅니다.

이 지표에서 DRL 기반 방법이 비슷한 성능을 보였습니다. 또한 모든 평가 지표에서 DRL 기반 방법이 휴리스틱 방법을 크게 앞섰습니다. 이러한 장점은 광범위한 훈련 샘플에서 패킹 패턴과 규칙성을 추출하는 DRL 기반 방법의 능력에 기인합니다. 반면 휴리스틱 방법은 특정 규칙이나 전략을 넘어 일반화하는 데 어려움을 겪을 수 있습니다. 기준선과의 비교는 우리 방법의 효율성을 나타냅니다. 또한, 그림 4에는 다양한 방법의 시각화된 패킹 결과를 정성적으로 비교한 결과가 나와 있습니다. 우리의 결과가 다른 경쟁 방법보다 시각적으로 우수하다는 것을 알 수 있습니다.

### C. 일반화

학습 기반 방법이 다양한 데이터 세트와 보이지 않는 시나리오에 걸쳐 일반화할 수 있는 능력은 지속적으로 조사와 관심의 대상이 되어 왔습니다. 이 섹션에서는 다양한 차원과 보이지 않는 항목의 다양한 빈에 대한 방법의 일반화 성능을 평가합니다.

*다양한 빈에 대한 일반화:* 앞서 언급한 학습을 위한 초기 빈 크기 외에도 , 빈 크기가 각각 30× 30× 30, 50× 50× 50, 100× 100× 100으로 설정된 세 가지 다른 환경을 소개합니다.
에 따라 데이터 세트의 항목 차원이 확장됩니다. 이러한 환경의 이름은 Bin-10, Bin-30, Bin-50 및 Bin-100입니다. 빈의 차원이 커짐에 따라 작업의 검색 공간이 증가하여 복잡성이 높아집니다.

표 III
$RS_{(sub)}$ 및 보이지 않는 항목이 포함된 두 개의 데이터 세트에서 평가할 때 $RS_{sub}$에 대해 학습된 정책의 성능

| Method | $RS_{sub}$ | | RS | | $RS_{exc}$ | |
|---|---|---|---|---|---|---|
| | *Uti* | *Num* | *Uti* | *Num* | *Uti* | *Num* |
| PCT [8] | 73.9% | 28.0 | 73.7% | 28.2 | 73.7% | 29.3 |
| Xiong et al. [3] | 73.8% | 27.9 | 73.0% | 27.8 | 72.9% | 29.0 |
| **GOPT** | **75.5%** | **28.7** | **76.1%** | **29.5** | **75.7%** | **30.2** |

를 사용하여 해결책을 찾습니다. 빈 차원에 대한 방법의 일반화 능력을 평가하기 위해, 저희는 Bin-10에서만 학습된 정책을 미세 조정 없이 다른 세 가지 환경으로 직접 전송합니다. 또한 설득력을 높이기 위해 여러 DRL 기반 기준 방법 [2], [3][8], , [10]과 함께 제안한 GOPT를 여러 환경에서 개별적으로 훈련하고 테스트합니다. Uti와 *Num*의 결과는 표 II에 요약되어 있습니다. Zhao 등의 방법 [2]은 Bin-100에서 수렴하지 못한다는 점에 주목할 필요가 있습니다. 표 II에 따르면 GOPT는 다양한 환경에서 일관된 성능을 유지할 뿐만 아니라 다른 방법보다 일관되게 우수한 성능을 보입니다. 중요한 것은 재학습을 하지 않는 정책 GOPT(Bin-10)가 학습 환경과 다른 환경에서도 안정적인 성능을 보인다는 점입니다. 다른 DRL 기반 방법은 다양한 빈 차원에 직면할 때 재학습이 필요하기 때문에 이러한 기능이 없습니다. 흥미롭게도 이들 중 일부는 Bin-30에서 상대적으로 높은 성능을 달성합니다. 이는 모델 매개변수의 증가와 이 크기에서 적당한 문제 복잡성 사이의 균형으로 인해 더 큰 빈에서 관찰되는 과도한 어려움 없이 향상된 피팅 능력을 발휘할 수 있기 때문인 것으로 추측됩니다. *보이지 않는 항목에 대한 일반화:* 또한, Bin-10에서 보이지 않는 항목을 사용하여 방법의 일반화 성능을 평가하는 실험을 수행합니다. 이 테스트는 모델이 다른 특성을 가진 테스트 데이터에 직면했을 때 성능이 저하되는 경우가 많기 때문에 매우 중요하고 까다로운 테스트입니다. 앞서 언급했듯이 RS 데이터 세트에는 125가지 유형의 항목이 있습니다. RS에서 25가지 유형의 항목($RS_{exc}$)을 무작위로 제외하여 하위 데이터 세트인 $RS_{sub}$로 에이전트를 훈련하고 전체 RS 및 $RS_{exc}$로 테스트합니다. 이전 실험에서 좋은 성과를 거둔 두 가지 기준선을 선택하여 비교합니다. 표 III에서 볼 수 있듯이, 하위 데이터셋에서 학습된 정책이 전체 데이터셋 RS와 보이지 않는 항목으로만 구성된 데이터셋 $RS_{exc}$모두에서 테스트했을 때 다른 정책보다 더 나은 성능을 보였습니다. 이는 다음을 시사합니다.

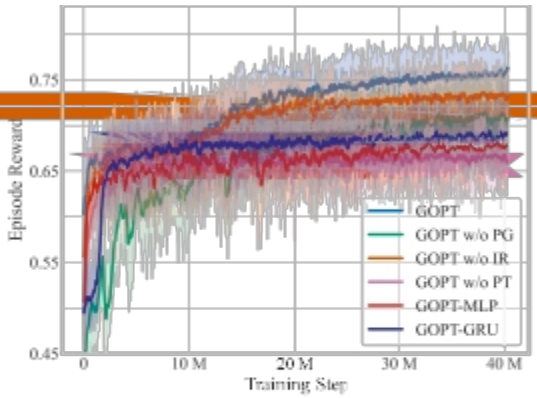



그림 5. 절제 연구에 대한 훈련 성능 비교. 결과는 128개의 서로 다른 무작위 시드를 사용하여 얻었습니다.

훈련된 정책은 보이지 않는 항목에 대해서도 적절한 일반화 능력을 보여줍니다. 또한 모든 항목에서 *Num*이 증가하는 것을 관찰했습니다.

메서드에 대한 RS 및 RS $_{exc}$ 이러한 데이터 세트에는 더 작고 포장하기 쉬운 품목을 늘릴 수 있습니다.

표 IV
다양한 보상 기능 비교

| Reward designs | *Uti* | *Num* |
|---|---|---|
| Step-wise | **76.1%** | **29.6** |
| Terminal [31] | 70.9% | 27.6 |
| Heuristic [9] | 72.4% | 28.0 |


그림 6. 실제 실험. (a) 로봇 포장 설정: KUKA 로봇에는 흡입 컵과 3개의 3D 카메라가 장착되어 있습니다. (b) 실패 사례: 로봇이 실험에서 실패의 주요 원인은 측정 오류이며, (c) 그리고 (d)는 안전한 포장과 단단한 포장의 스냅샷입니다.

### *D. 절제 연구*

다양한 구성 요소의 영향을 철저하게 분석하기 위해 추가 제거 연구를 수행합니다. 이러한 구성 요소에는 배치 생성기(PG), 아이템 표현(IR), 패킹 트랜스포머(PT)가 포함됩니다. 저희는 PG를 제외하고 모든 배치와 해당 마스크를 신경망에 제공하여 그 효과를 설명합니다. 또한 아이템 표현을 3차원 벡터에서 제안한 모드로 변환하지 않고 얻은 결과도 제시합니다. 또한 PT를 제거하고(GOPT w/o PT), PT를 MLP(GOPT-MLP) 및 GRU(GOPT-GRU)로 대체하여 실험을 수행하여 그 중요성에 대한 인사이트를 얻었습니다. 결과는 표 I에 나와 있습니다. 또한 보상 곡선과 훈련 단계를 그림 5에 제시했습니다.

표 1과 그림 5에서 볼 수 있듯이, 이 연구에서 도입된 세 가지 구성 요소는 모두 기대에 부합하는 양호한 결과를 보여줍니다. 비교 분석 결과, PT를 사용하지 않은 GOPT, GOPT-MLP, GOPT-GRU는 GOPT에 비해 성능이 현저히 저하된 것으로 나타났습니다. 이는 성능 향상에 있어 제안된 PT를 통한 공간 관계 식별의 이점을 강조합니다. 이러한 능력은 다른 네트워크에 비해 복잡한 순차적 데이터를 처리하고 장거리 종속성을 학습하는 데 있어 주의 메커니즘의 탁월한 효율성에 기인합니다. 또한 그림 5에서 볼 수 있듯이, PT를 통합한 모델(GOPT, GOPT w/o PG, GOPT w/o IR)은 PT가 없는 모델(GOPT w/o PT, GOPT-MLP, GOPT-GRU)보다 약 3천만 대 1천만 개의 훈련 데이터가 더 많이 필요해 컨버전스를 달성하는 데 더 많은 시간이 걸립니다. 또한 GOPT는 IR이 없는 GOPT보다 공간 활용도가 높고 더 많은 항목을 포함하므로 제안된 항목 표현이 DRL 에이전트의 학습과 최종 성능을 촉진한다는 것을 나타냅니다. 그림 5를 보면 PG를 사용하지 않는 GOPT가 훈련 초기 단계에서 가장 적은 보상을 얻는다는 것을 알 수 있습니다. 이는 인간의 경험에 의해 정보를 얻는 PG 모듈이 다음을 개선하는 데 기여할 수 있음을 시사합니다. DRL 에이전트가 아직 상당한 포장 지식을 축적하지 못한 경우 샘플링 효율성을 높일 수 있습니다.

또한 본 연구에서 사용한 단계별 보상, 에피소드의 최종 공간 활용으로 정의되는 터미널 보상[31], 불합리한 행동으로 인한 공간 낭비를 방지하기 위해 패널티 조건을 추가한 휴리스틱 보상[9]을 포함하여 확률 문제에 대한 보상 설계의 영향을 조사합니다. 표 IV에 따르면, 터미널 보상으로 훈련된 에이전트가 가장 낮은 성능을 보인 반면, 단계별 보상은 휴리스틱 보상보다 단순하고 직관적임에도 불구하고 더 효율적인 것으로 나타났습니다.

### *E. 실제 실험*

우리는 그림 6(a)와 같이 실제 로봇 포장 테스트베드를 구축하여 실제 환경에서 우리 방법의 적용 가능성을 검증합니다. 물품 포장용 빈의 크기는 56cm× 36.5cm× 21cm이며, 각 셀의 길이는 0.7cm인 80× 52× 30개의 빈으로 이산화됩니다. 이 작업에서 로봇은 상자에서 상자를 선택하고 Lucid 카메라의 시야 내에서 이동하여 상자의 크기와 손에 든 자세를 평가한 다음 시뮬레이션에서 학습한 GOPT에 따라 다른 상자에 넣습니다. 한편 두 대의 카메라가 장착되어 이 두 상자를 별도로 모니터링합니다. 패킹 빈의 하이트맵은 포인트 클라우드의 분할 및 투영과 직사각형 감지를 통해 생성됩니다. 피킹할 상자가 남지 않거나 다음 상자를 포장할 공간이 부족할 때까지 픽 앤 팩 프로세스가 진행됩니다. 실제 시나리오에서 로봇이 이 방법을 활용하여 포장 작업을 완료할 수 있음을 보여줍니다. 데모 동영상은 보충 자료에서 확인할 수 있습니다.

실험을 통해 카메라로 인한 측정 오류가 박스 간 충돌을 일으킬 가능성이 있음을 관찰했습니다.

이 발생할 수 있습니다(그림 6(b) 참조). 이를 방지하기 위해 추가 그림 6(c)와 같이 배치된 각 상자 주변에 0.7cm의 버퍼 공간이 할당되어 20개의 테스트에서 평균 공간 활용률이 67.5%로 나타납니다. 버퍼를 0으로 줄이면 오류 위험이 증가하고 20개 테스트 중 2개가 실패하지만 그림 6(d)에서 볼 수 있듯이 사용률이 73.3%(18개 테스트 성공 시)로 더 높아집니다. 이러한 결과는 실제 로봇 포장 시나리오에서 시스템 신뢰성과 포장 결과물의 소형화를 모두 향상시키기 위한 향후 연구 에 원동력을 제공합니다.

## V. 결론

저희는 온라인 3D 빈 패킹을 위해 GOPT라는 새로운 프레임워크를 제공합니다. GOPT는 배치 후보를 생성하고 이러한 후보가 있는 빈의 상태를 표현하기 위해 배치 생성기 모듈을 사용합니다. 한편 패킹 트랜스포머( )는 물품과 빈의 정보를 효과적으로 융합하는 교차 주의 메커니즘을 사용하여 패킹의 공간적 상관관계를 식별합니다. 광범위한 실험을 통해 GOPT가 기존 방법보다 우선순위가 높다는 것이 입증되었으며, 포장 성능뿐만 아니라 일반화 기능에서도 현저한 향상( )을 보여주었습니다. 특히, 훈련된 GOPT 정책은 다양한 빈과 보이지 않는 품목 모두에 걸쳐 일반화할 수 있습니다. 마지막으로 학습된 패킹 정책을 로봇 시스템( )에 성공적으로 적용하여 실제 적용 가능성을 입증했습니다. 향후에는 로봇 픽 앤 플레이스 작업에서 흔히 발생하는 문제인 불규칙한 모양의 물체를 포장하는 데까지 이 방법을 확대 적용할 계획입니다. 또한 실제 로봇 포장 시스템의 신뢰성을 개선하는 방법도 모색할 계획입니다.

## 참조


[1] F. Wang과 K. Hauser, "불규칙하고 새로운 3D 물체의 고밀도 로봇 패킹," *IEEE Trans. Robot.*, 38권, 2호, 1160-1173쪽, 2022년 4월.

[2] H. Zhao, Q. She, C. Zhu, Y. Yang 및 K. Xu, "온라인 3D 빈 패킹 with constrained deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, 35, 1, 741-749쪽.

[3] H. Xiong, K. Ding, W. Ding, J. Peng 및 J. Xu, "딥 강화 학습에 기반한 신뢰할 수 있는 로봇 포장 시스템을 향하여", *Adv. Inform.*, vol. 57, 2023, 예술 번호 102028.

[4] O. X. d. Nascimento, T. A. d. Queiroz 및 L. Junqueira, "컨테이너 적재 문제의 실용적인 제약: 포괄적인 공식- tions 및 정확한 알고리즘", *Comput. Operations Res.*, vol. 128, 2021, 예술 번호 105186.

[5] C. Ha, T. T. 응우옌, L. T. 부이, R. 왕, "동적 환경과 물리적 인터넷에서 3차원 컨테이너 적재 문제에 대한 온라인 패킹 휴리스틱," in *Proc. Eur. Conf. Appl. Evol. Computation*, 2017, 140-155쪽.

[6] R. Verma 외, "온라인 3D 빈 패킹을 위한 일반화된 강화 학습 알고리즘", 2020, *arXiv:2007.00463*.

[7] Z. Yang et al., "PackerBot: 휴리스틱을 이용한 가변 크기 제품 포장 심층 강화 학습," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, 5002-5008쪽.

[8] H. Zhao, Y. Yu, and K. Xu, " 패킹 구성 트리에서 효율적인 온라인 3D 빈 패킹 학습," in *Proc. Conf. Learn. Representations*, 2022.

[9] S. Yang et al., " 온라인 3D 빈 패킹을 위한 휴리스틱 통합 심층 강화 학습," *IEEE Trans. Automat. Sci. Eng.*, vol. 21, no. 1, pp. 939-950, 2024년 1월.

[10] H. Zhao, C. Zhu, X. Xu, H. Huang 및 K. Xu, "온라인 3D 빈 패킹을 위한 실질적으로 실현 가능한 정책 학습," *Sci. China Inf. Sci.*, 65, 1, 1-17, 2022, pp.

[11] S. 알리, A. G. 라모스, M. A. 카라비야, J. F. 올리베이라, "온라인 3차원 패킹 문제: 오프라인 및 온라인 솔루션 접근 방식에 대한 검토," *Comput. Ind. Eng.*, vol. 168, 2022, 예술 번호 108122.

[12] G. 도사 및 J. 스갈, "퍼스트 핏 빈 패킹: 긴밀한 분석", *Proc. Symp. Theor. Aspects Comput. Sci. (2013). Schloss Dagstuhl-Leibniz- Zentrum Fuer Informatik*, 2013.

[13] G. Dósa 및 J. Sgall, "최적 빈 포장의 최적 분석", *Proc. Colloq. Automata, Languages, Program.*, 2014, 429-441쪽.

[14] L. Wang, S. Guo, S. Chen, W. Zhu 및 A. Lim, "실제 로딩 제약 조건이 있는 3D 패킹을 위한 두 가지 자연 휴리스틱," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2010, 256-267쪽.

[15] S. 마르텔로, D. 피징거, D. 비고, "3차원 빈 포장 문제", *Operations Res.*, 48권 2호, 256-267쪽, 2000.

[16] T. G. Crainic, G. Perboli 및 R. Tadei, "3차원 빈 패킹을 위한 극단적인 포인트 기반 휴리스틱," *Informs J. Comput.*, 20권, no. 3, pp. 368-384, 2008.

[17] F. Parreño, R. Alvarez-Valdés, J. M. Tamarit, J. F. Oliveira, "컨테이너 로딩 문제에 대한 최대 공간 알고리즘," *INFORMS. J. Comput.*, vol. 3, 412-422쪽, 2008.

[18] M. Agarwal, S. Biswas, C. Sarkar, S. Paul, 및 H. S. Paul, "Jampacker: 직육면체 물체를 위한 효율적이고 안정적인 로봇 빈 포장 시스템," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 319-326, Apr.

[19] A. 야림캠, S. 아스타, E. *외 즈 칸*, A. 제이 파크스, "온라인 빈 패킹을 위한 파라미터 조정을 통한 휴리스틱 생성", *Proc. IEEE Symp. Learn. Syst.*, 2014, 102-108쪽.

[20] M. 로페즈-이바네즈, J. 두부아-라코스테, L. P. 카세레스, M. 비라타리, 및 T. Stützle, "irace 패키지: 자동 알고리즘 구성을 위한 반복 레이싱," *Operations Res. Perspectives*, 3권, 43-58페이지, 2016.

[21] F. Wang과 K. Hauser, "로봇 조작기를 사용한 비볼록 3D 물체의 안정적인 빈 패킹", *Proc. Conf. Robot. Automat.*, 2019, 8698- 8704쪽.

[22] W. Shuai 외, "불가피한 불확실성이 있는 규정 준수 기반 로봇 3D 빈 포장", *IET 제어 이론 응용*, 17권, 2241-2258쪽, 2023.

[23] R. Hu, J. Xu, B. Chen, M. Gong, H. Zhang, 및 H. Huang, "TAP-Net: 강화 학습을 이용한 수송 및 포장," *ACM Trans. Graph.*, 39권 6호, 1-15쪽, 2020.

[24] W. Kool, H. v. Hoof 및 M. Welling, "주의, 라우팅 문제 해결 방법 배우기!", in *Proc. Conf. Learn. Representations*, 2019. [온라인]. 사용 가능: https://openreview.net/forum?id=ByxBFsRqYm

[25] M. Nazari, A. Oroojlooy, L. Snyder 및 M. Takác, "차량 라우팅 문제 해결을 위한 강화 학습", *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018,vol.

[26] Q. Que, F. Yang, 및 D. Zhang, "변압기 네트워크와 강화 학습을 이용한 3D 패킹 문제 해결," *Expert Syst. Appl.*, vol. 214, 2023, 예술 번호 119153.

[27] O. 쿤두, S. 두타, S. 쿠마르, "Deep-pack: 심층 강화 학습을 이용한 비전 기반 2D 온라인 빈 패킹 알고리즘," in *Proc. Conf. Robot Hum. Interactive Commun. (RO-MAN)*, 2019, pp. 1-7.

[28] Y. 우, E. 만시모프, R. B. 그로스, S. 리아오, J. 바, "크로네커 계수 근사법을 사용한 심층 강화 학습을 위한 확장 가능한 신뢰 영역 방법," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017,vol.

[29] P. Velic˘kovic´, G. 쿠쿠룰, A. 카사노바, A. 로메로, P. 리오, Y. 벤지오, "그래프 주의 네트워크", 2017, *arXiv:1710.10903*.

[30] V. Mnih 외, "심층 강화 학습을 위한 비동기 방법", *Proc. Conf. Mach. Learn.*, 2016, 1928-1937쪽.

[31] J. Xu, M. Gong, H. Zhang, H. Huang 및 R. Hu, "Neural packing: 시각적 감지부터 강화 학습까지," *ACM Trans. Graph.*, 42권, 6호, 1-11쪽, 2023.

[32] P. Li 외, "SelfDoc: 자기 감독형 문서 표현 학습," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, 5648-5656쪽.

[33] J. 슐만, F. 울스키, P. 다리왈, A. 래드포드, O. 클리모프, "근거리 정책 최적화 알고리즘", 2017, *arXiv:1707.06347*.

[34] J. 슐만, P. 모리츠, S. 레빈, M. 조던, P. 아벨, "일반화된 이점 추정을 이용한 고차원 연속 제어", 2015, *arXiv:1506.02438*.

[35] J. Weng 외, "Tianshou: 고도로 모듈화된 심층 강화 학습 라이브러리," *J. Mach. Learn. Res.*, 23, 1, 12275-12280, 2022.