Unveiling the Significance of Toddler-Inspired Reward Transition in Goal-Oriented Reinforcement Learning

Junseok Park[1], Yoonsung Kim[1*], Hee Bin Yoo[1*], Min Whoo Lee[1], Kibeom Kim[1], Won-Seok Choi[1], Minsu Lee[1,2†], Byoung-Tak Zhang[1,2,†]

[1]Seoul National University
[2]AI Institute of Seoul National University (AIIS)
{jspark, yskim, hbyoo, mwlee, kbkim, wchoi, mslee, btzhang}@bi.snu.ac.kr

Abstract

Toddlers evolve from free exploration with sparse feedback to exploiting prior experiences for goal-directed learning with denser rewards. Drawing inspiration from this **Toddler-Inspired Reward Transition**, we set out to explore the implications of varying reward transitions when incorporated into Reinforcement Learning (RL) tasks. Central to our inquiry is the transition from sparse to potential-based dense rewards, which share optimal strategies regardless of reward changes. Through various experiments, including those in egocentric navigation and robotic arm manipulation tasks, we found that proper reward transitions significantly influence sample efficiency and success rates. Of particular note is the efficacy of the toddler-inspired Sparse-to-Dense (S2D) transition. Beyond these performance metrics, using *Cross-Density Visualizer* technique, we observed that transitions, especially the S2D, smooth the policy loss landscape, promoting wide minima that enhance generalization in RL models.

Introduction

In early years, toddlers behave much like exploratory agents. Throughout their development, they interact with their surroundings without much prior knowledge, akin to someone embarking on new experiences without expecting immediate rewards (Oudeyer and Smith 2016). As they grow, toddlers transition from free exploration to more goal-directed learning, aiming for specific goals, resembling someone working towards known rewards for their efforts (Gopnik, Meltzoff, and Kuhl 1999; Gibson 1988; Piaget, Cook et al. 1952; Gopnik et al. 2017) as illustrated in Figure 1.

This learning pattern in toddlers can be incorporated in Reinforcement Learning (RL), as illustrated in Figure 1-(a). In RL, agents learn by interacting with their environment and receiving feedback, much like how toddlers learn from their interactions. Similar to toddlers, agents must navigate towards positive feedback they receive, which can be infrequent (sparse) or detailed (dense). Sparse feedback might mean that the agent requires more attempts to figure out the desired behavior due to limited guidelines (Andrychowicz et al. 2020; Knox et al. 2023). Meanwhile, dense feedback can guide the agent faster but might inadvertently focus them on immediate outcomes, missing out on the bigger picture or long-term strategies (Laud 2004).

Given these intricacies, simply sticking to one type of feedback might not capture the essence of learning. Drawing inspiration from toddler developmental stages, blending both feedback types could provide richer insights. The transition toddlers make from free exploration – akin to sparse feedback – to specific, goal-driven learning – similar to dense feedback in RL – offers a unique perspective, as shown in Figure 1-(a). With this perspective, our paper addresses the following question: *"How does reward transition that proceeds in a sparse-to-dense manner, inspired by toddler learning, affect the learning of agents?"* Through a series of experiments, including egocentric navigation and robotic arm manipulation tasks, we aim to explore the **Toddler-Inspired Sparse to Potential-based Dense (S2D) Reward Transition**. Our goal is not only to explore its efficacy but also to delve into the underpinning reasons by analyzing its comparative advantages against other rewards or prevalent strategies in the field of RL.

Taking the concept of "reward transition in learning" further, we consider visualizing the learning parameters as a topographical map. One type of such landscape representation, a policy loss landscape provides an intuitive visualization of learning dynamics in RL (Li et al. 2018). On this map, each point corresponds to a set of learning parameters, and its altitude signifies the loss value. As in any landscape, some areas are rugged, featuring steep mountains or deep valleys, indicating challenging learning regions. Understanding the notion of smoothness of loss landscape is vital in neural network-based learning. Smooth terrains in this landscape, devoid of abrupt pitfalls, enable quicker and more reliable convergence through gradient descent.

Critically, smoother landscapes often promote wide minima, which are associated with better generalization of learned policies to novel situations in dynamic environments (Keskar et al. 2017). Empirically, we observed that employing the Sparse-to-Dense (S2D) Reward Transition made this terrain smoother by reducing the depth of local minima, as illustrated in Figure 1-(b).

Our study contributes to a deeper understanding of the intricate balance between exploration and exploitation and provides insight into designing reward structures in RL.

[*]Equal Contributions.
[†]Corresponding Authors.

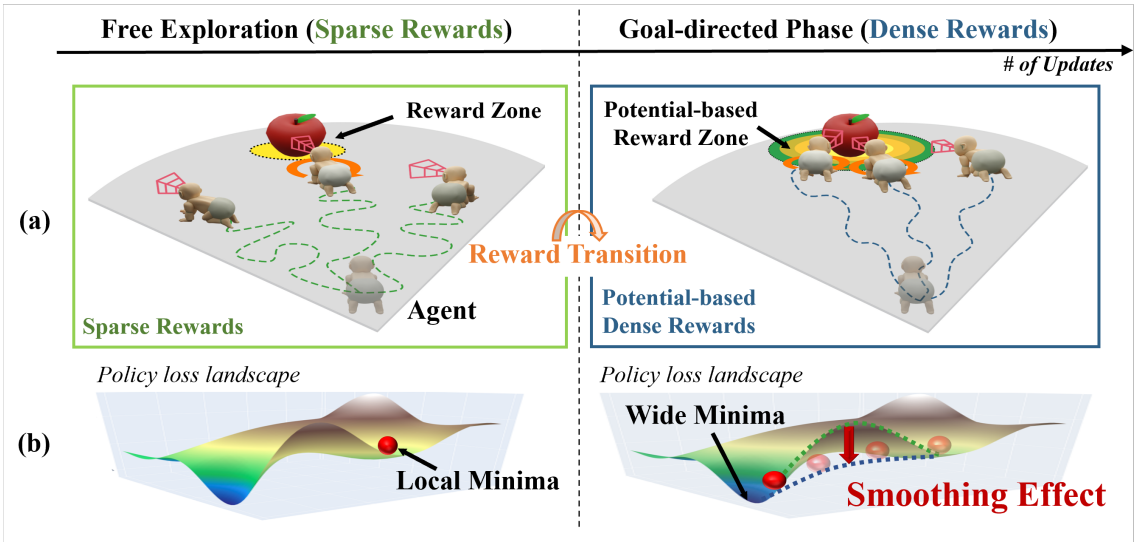


Figure 1: Parallel learning trajectories: toddlers and agents. (a) The figure compares a toddler's learning journey with an agent's. On the left, a toddler freely explores the apple, symbolizing sparse reward learning. As we transition right, the toddler's focus on specific tasks reflects goal-guided learning. Similarly, the agent's progression from sparse to potential-based dense rewards is charted above, highlighting parallels in learning evolution. (b) As reward transitions occur, the depth of local minima reduces, leading to a wide minima via the smoothing effect, thereby enhancing more generalization.

By emulating the learning processes observed in toddlers, we hope to bridge the gap between biological and artificial learning mechanisms. Based on this, we have offered a novel perspective in applying and addressing this synthesis for more robust, adaptable, and efficient RL systems.

The main contributions of this paper can be summarized as follows: **(1)** We observed that the Toddler-Inspired Reward Transition enhances success rates, sample efficiency, and generalization within goal-conditioned RL. **(2)** Our experimental analyses support that such transition has smoothing effects on the policy loss landscape and promotes wide minima, corroborating the performance improvements. **(3)** Our findings highlight the potential of biologically inspired approaches in providing clues for exploration-exploitation tradeoff and reward shaping challenges.

Related Work

Toddler-inspired learning. Drawing insights from toddler developmental stages has provided a fresh perspective in advancing deep learning. By harnessing the innate exploratory tendencies and distinctive learning mechanisms of toddlers, researchers have refined both supervised and reinforcement learning techniques. For instance, classifiers trained on datasets of toddlers' perspectives on objects outperformed those using adults' perspectives (Bambach et al. 2018), underscoring the potential of leveraging toddlers' exploration mechanism. Similarly, critical learning phases in toddlers have counterparts in RL (Park et al. 2021; De Kleijn, Sen, and Kachergis 2022) and deep networks (Achille, Rovere, and Soatto 2018). Such toddler-inspired methodologies emphasize the alignment between toddler growth and AI model evolution, underscoring the potential of biological insights in driving AI forward.

Exploration-exploitation in deep RL. Balancing exploration with exploitation is an inherent challenge of RL (Ladosz et al. 2022). Moreover, deep RL intensifies this complexity with high-dimensional spaces, like raw image inputs, where pixels are unique dimensions. To mitigate this, many algorithms favor exploration, utilizing tools like intrinsic motivation (Badia et al. 2020; Aubret, Matignon, and Hassas 2019; Pathak et al. 2017). Drawing from toddlers' learning, we observe the transition from free exploration to specific exploitation based on collected previous experience. Our approach offers a fresh take on RL's exploration-exploitation dilemma without introducing algorithmic complexities.

Curriculum learning. Curriculum Learning (CL), inspired by curricula in human's education, has been known to boost performance, training speed (Hacohen and Weinshall 2019), and safety (Turchetta et al. 2020) in machine learning. CL's progression from simple to complex tasks promotes generalization and convergence (Bengio et al. 2009; Weinshall, Cohen, and Amir 2018) in both supervised and reinforcement learning (Florensa et al. 2018; Graves et al. 2017; Narvekar and Stone 2020). Unlike the studies that initially restrict the diversity to easy tasks (Kalantidis et al. 2020; Du et al. 2021; Dong, Gong, and Zhu 2017), several studies (Zhang 1994; MacKay 1992) advocate a *general-to-specific* approach, where agent initially collects varied learning experiences and then exploits these experiences later in the curriculum. Adapting the S2D transition observed in toddlers into RL, we embrace this philosophy in reward transition for goal-oriented tasks.

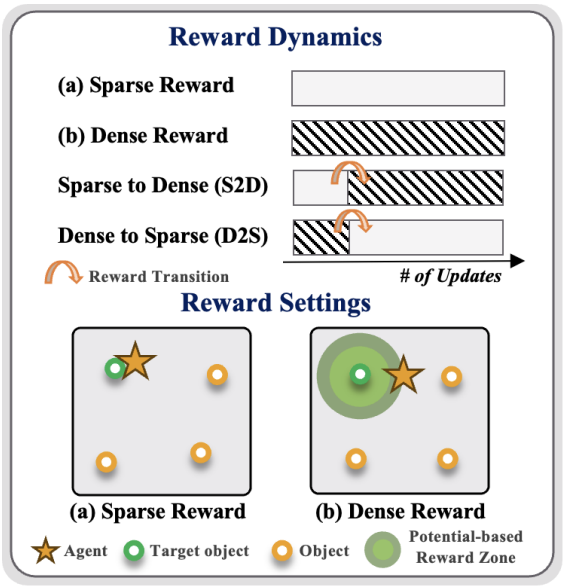


Figure 2: Overview of the baseline rewards. The S2D presents reward inspired by toddler learning. In sparse rewards, agents are rewarded upon reaching the target. For potential-based dense rewards, they get an extra reward determined by the distance to a specific unit from the object.

Potential-based reward shaping (PBRS). In RL, maximizing cumulative rewards guides agent behavior. However, due to the inherent challenges in designing optimal reward functions for various tasks, it often necessitates a process known as reward engineering. Reward Shaping (RS) is one such technique that enhances training by supplementing the environment's feedback (Taylor and Stone 2009). Particularly, in environments where the reward changes, an additional reward from a *potential function* is employed to ensure the agent's optimal strategy remains unaffected (Ng, Harada, and Russell 1999). Commonly, such shaped rewards are consistently applied throughout training. Unlike this, we explore the **Toddler-Inspired Reward Transition**, focusing on the effects of changing the density of rewards.

Preliminaries

Reinforcement learning. RL is a field of machine learning in which the agent learns through trial and error, similar to how humans acquire skills. It is applied to various tasks that involve sequential decision making. A widely used formulation of RL problem, Markov Decision Process (MDP) is defined as $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is a set of environment states, $\mathcal{A}$ is a set of possible actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition probability distribution, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function and γ is a discount factor. At every time step $t \in \mathbb{N}$, the agent in the current state $s_t \in \mathcal{S}$ performs the action $a_t \in \mathcal{A}$ according to a policy $\pi(\cdot|s_t)$, and receives the next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ and reward $\mathcal{R}(s_t, a_t)$. RL algorithms aim to obtain an optimal policy $\pi^* \in \Pi^*$ that maximizes the expected cumulative rewards $R = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}\left(s_t, a_t\right)\right]$ with γ applied, where Π^* is the set of optimal policies.

Curriculum learning. Curriculum Learning (CL) is a strategy to train an ML model using tasks that gradually increase in difficulty. In RL, CL can be formulated as a learning framework where the agent is trained on a sequentially changing series of MDPs $\mathcal{M}_i = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_i, \gamma \rangle$.

Definition 1 (Curriculum) *Let $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_N$ be a sequence of MDPs $\mathcal{M}_i = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_i, \gamma \rangle$, and $\mathcal{T} = (T_1, T_2, \cdots, T_{N-1}) \in \mathbb{N}^{N-1}$. A curriculum is a tuple $\mathscr{C} = (\{\mathcal{M}_i\}_{i=1}^N, \mathcal{T})$ where the agent is trained on $\mathcal{M}_{I(t;\mathcal{T})}$ at training step t. Stage indicator $I(t;\mathcal{T})$ is defined as:*

$$I(t;\mathcal{T}) := i, \quad t \in [T_{i-1}, T_i)$$

for each stage $i \in \{1, \cdots, N\}$, where $T_0 := 0$ and $T_N := \infty$. We call $\mathcal{T} = (T_1, T_2, \cdots, T_{N-1})$ the stage transitions.

CL also boosts the training by arranging the tasks as *"general-to-specific,"* where the agent is provided rewards with monotonically increasing densities over training. Formally, we say an environment is sparse when only a small portion of the state space is included in $\text{supp}(\mathcal{R})$. That is,

$$|\text{supp}(\mathcal{R})| \ll |\mathcal{S}|, \quad \text{where}$$
$$\text{supp}(\mathcal{R}) = \{s \in \mathcal{S} \mid \exists a \in \mathcal{A} \ \ s.t. \ \ \mathcal{R}(s,a) \neq 0\}.$$

$\text{supp}(\mathcal{R})$ means the region *supported* by reward function $\mathcal{R}$ which has non-zero reward for some actions.

Wide minima phenomenon and loss landscape. Deep neural networks traverse a high-dimensional *loss landscape*, with altitude indicating the loss for specific parameters (Li et al. 2018). The aim is to find the *minima*. In *wide minima*, due to broad gradients, gradient descent is more likely to converge smoothly to global minima. This fosters robustness and superior generalization to new data (Keskar et al. 2016). Conversely, in *sharp minima*, steep gradients can trap models in local minima, resulting in overfitting and poor generalization across diverse data distributions (Goodfellow, Vinyals, and Saxe 2014). Empirically, models within wide minima demonstrate better performance and generalization than those in sharp minima (Keskar et al. 2017; Jastrzębski et al. 2018). In deep RL as well, where the distribution of agent's experiences may slightly vary every time step, policies in wide minima could improve in generalization.

Toddler-Inspired Reward Transition

To conduct our experiments, we need to formulate and design a toddler-inspired reward transition in RL, emulating the toddler reward transition paradigm. Furthermore, we analyze the influence of this reward transition on the learning behavior of agents, focusing on the policy loss landscape and the wide minima phenomenon.

Toddler-Inspired Sparse to Potential-based Dense Reward Curriculum

In this paper, we harness curriculum learning from the perspective of encouraging exploration-to-exploitation as a Sparse to potential-based Dense (S2D) reward transition curriculum and explain how this learning mechanism can benefit RL. We define $\mathscr{C} = (\{\mathcal{M}_i\}_{i=1}^N, \mathcal{T})$ as an *S2D-curriculum*

if reward functions of MDPs $(\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_N)$ become progressively denser while preserving optimality.

Definition 2 (***Toddler-inspired S2D-curriculum***) *A curriculum* $\mathscr{C} = (\{\mathcal{M}_i\}_{i=1}^N, \mathcal{T})$ *with MDPs* $\{\mathcal{M}_i\}_{i=1}^N$ *is an S2D-curriculum if following conditions are satisfied:*

$$\text{supp}(\mathcal{R}_1) \subseteq \text{supp}(\mathcal{R}_2) \subseteq \cdots \subseteq \text{supp}(\mathcal{R}_N) \quad (1)$$

$$\Pi_1^* \supseteq \Pi_2^* \supseteq \cdots \supseteq \Pi_N^*, \quad (2)$$

Π_i^* *is a set of optimal policies with MDP* $\mathcal{M}_i$*. Here, we call* $\{\mathcal{R}_i\}_{i=1}^N$ *a guidance of the curriculum* $\mathscr{C}$*.*

The first condition in Equation 1 denotes that the reward function becomes denser, *i.e.*, the guidance becomes more explicit. The second condition in Equation 2 constrains the optimality to be preserved during the transition of the MDPs, *i.e.*, the optimal policies of $\mathcal{M}_i$ are also optimal in $\mathcal{M}_{i+1}$. At a high level, the sequence of MDPs in the above definition is *S2D-curricular* in the sense that the reward functions are arranged in a *"sparse-to-dense"* order.

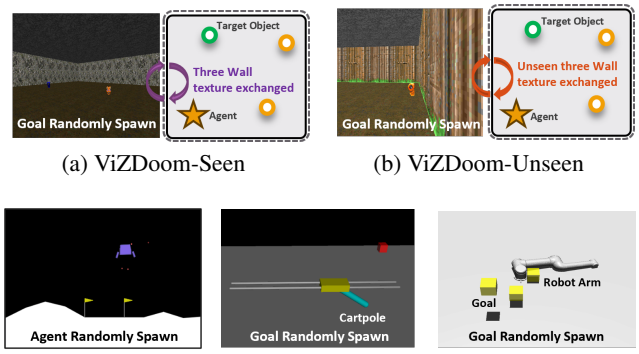
Visualizing Post-Transition 3D Policy Loss Landscape: Cross-Density Visualizer

As seen in Figure 6 and Appendix B, our study delves into the effect of the S2D transition on policy loss landscape, reflecting toddlers' cognitive evolution (Gopnik et al. 2017; Piaget, Cook et al. 1952). Following (Li et al. 2018), we visualize the policy loss landscapes using grids of parameters $\tilde{\theta} = \theta + \alpha\mathbf{x} + \beta\mathbf{y}$. Here, θ signifies current parameters and α, β are normalized coordinates. Vectors $\mathbf{x}$ and $\mathbf{y}$ that form the axes arise from two specific perturbations in the network's parameter space. $\mathbf{x}$ and $\mathbf{y}$ are made unit vectors by a normalized filter for consistent scaling and clarity. The Z-axis captures policy loss, averaged over a batch of transitions from the replay buffer. We are not concerned with the altitude and which landscape is above or below another, because the two landscapes correspond to separate network parameters, each of them having its own loss range according to its current learning progress.

Noting the lack of visualization techniques in prior studies for policy loss landscapes based on reward transitions, we design the Cross-Density Visualizer. This method portrays the 3D policy loss landscape during transitions from purely sparse or dense rewards to mixed-reward settings. Thus, one set includes Sparse-to-Dense (S2D) and Sparse-to-Sparse (Only Sparse), while the other contains Dense-to-Sparse (D2S) and Dense-to-Dense (Only Dense). Our representations, displayed in Figure 6 and expanded upon in Appendix B, highlight comparable *smoothing effects* particularly in the S2D model.

Exploring Minima Sharpness After Reward Transitions

Observing a reduction in the depth of local minima due to smoothing effects led us to hypothesize that the S2D transition promotes escape from local minima and enhances generalization in wide minima. Wide minima in neural networks can serve as a measure indicating robust and adaptable models (Keskar et al. 2017; Jastrzębski et al. 2018). By exploring



(a) ViZDoom-Seen (b) ViZDoom-Unseen

(c) LunarLander-V2 (d) CartPole-Reacher (e) UR5-Reacher

Figure 3: Experimental environments. In particular, (a) and (b) are environments for evaluating generalization.

minima with this transition, we aspire for both performance and a deeper grasp of agent adaptability in diverse scenarios. To check how much the policy resides in a wide minima, we measure the end-of-training convergence of S2D's neural network to wide minima using the sharpness metric of Equation 3 and compare with those of baselines in the same way as proposed in (Foret et al. 2021), which is a specific form of sharpness measure proposed in (Keskar et al. 2017).

$$\max_{||\epsilon||_2 \leq \rho} L_\pi(\theta + \epsilon) - L_\pi(\theta) \quad (3)$$

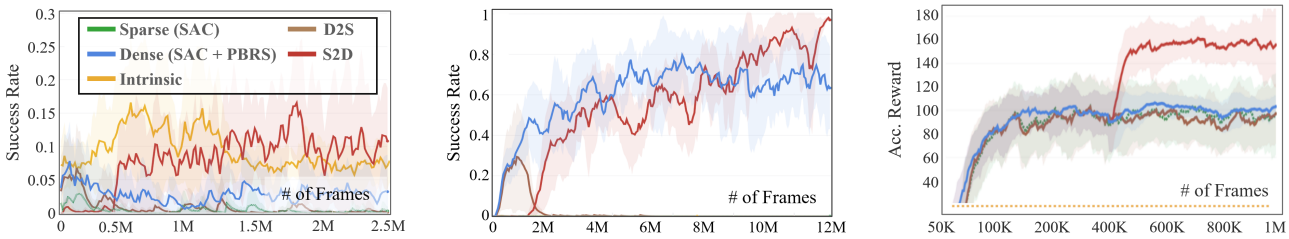
Here, θ is the current parameter in the policy loss landscape. The maximizer ϵ can be estimated from $\hat{\epsilon} = \rho\,\text{sgn}(\nabla_\theta L_\pi(\theta)) \cdot |\nabla_\theta L_\pi(\theta)|^{q-1} / \left(||\nabla_\theta L_\pi(\theta)||_q^q\right)^{\frac{1}{p}}$, where $1/p + 1/q = 1$, $\text{sgn}(\cdot)$ is element-wise sign function (Foret et al. 2021). We use $\rho = 0.02, p = 2$ in our experiments.

Experiments

In our experimental section, we delve deeply into the dynamics of the S2D reward transition compared to multiple reward-driven methods. We unveil its substantial effects within multiple challenging environments, which are illustrated in Appendix A. We particularly explored the implications of applying the reward transition to RL by probing three critical questions:

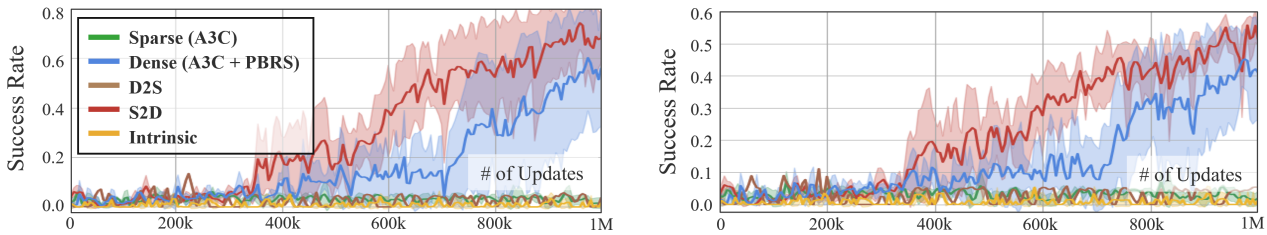
- **Performance Enhancement**: How does the Toddler-inspired S2D reward transition compare to other diverse reward settings?
- **Post-Transition 3D Policy Loss Landscape**: What are the effects of the S2D transition on the policy loss landscape?
- **Correlation Between Wide Minima and Toddler-Inspired Reward Transition**: Does the S2D transition foster convergence to wide minima?

Our understanding is further supported by extensive supplementary experiments in Appendix C.



(a) CartPole-Reacher (b) UR5-Reacher (c) LunarLander-V2

Figure 4: Performance of the agent with various reward types in multiple goal-oriented tasks. Notably, in (c) LunarLander, the accumulated reward from intrinsic rewards was significantly below zero, indicated by a dashed line.



(a) ViZDoom-Seen (b) ViZDoom-Unseen

Figure 5: Generalization performance of the ViZDoom agent with various types of rewards.

Reward Setting Details

Design of sparse and dense reward. In the sparse reward setting, the agent receives a reward only upon success, i.e., when reaching the goal. In the dense reward setting, the agent receives a potential-based reward based on its proximity to the goal. This is expressed as $\psi(s) = \text{diam}(\mathcal{S}) - ||s - g||_2$, where $\mathcal{S}$ is the set of states, $g \in \mathcal{S}$ is the goal state, and $\text{diam}(\cdot)$ denotes the diameter of the given set.

Reward-driven baselines for comparison. Aiming to investigate the relation between exploration-exploitation tradeoff and reward transitions, we explore the Sparse-to-Dense (S2D) approach to mirror toddlers' developmental progression. We also evaluate its counterpart, Dense-to-Sparse (D2S), and solely sparse or dense reward schemes for a comprehensive assessment of reward strategies. Given the prominence of intrinsic motivation in tackling exploration-exploitation, we adopted NGU (Badia et al. 2020)—designed for discrete environments like ViZDoom and LunarLander—and RND (Burda et al. 2018)—tailored for continuous action such as CartPole and UR5—as additional baselines. Both methods incentivize agents to explore by providing additional intrinsic rewards for discovering novel states.

Design for reward transition. We also explored the hyperparameter for reward transition through ablation studies as seen in Table 1. Recognizing the importance of the temporal aspects of initial cognitive and motor interactions in early developmental stages (Piaget, Cook et al. 1952; Shonkoff and Phillips 2000), we divide the first quarter and segmented this period into three points for reward transition. We set the transition timings at $t \in \{1N, 2N, 3N\}$, where N corresponds to roughly a third of the first quarter of the entire training duration. The specific value of N, adjusted for each environment's episode length, is detailed in the Appendix A. We denote these phases as $\mathscr{C}_1$, $\mathscr{C}_2$, and $\mathscr{C}_3$, signifying the S2D or D2S reward transition points.

Environment Details

We assess the efficacy of reward dynamics across diverse conditions, encompassing state and visual observations as well as both discrete and continuous action domains, as seen in Appendix A. We evaluate the reward settings, including S2D reward transition, across a variety of goal-based tasks in well-established benchmark environments. As seen in Figure 3, these include LunarLander (Brockman et al. 2016), CartPole, and UR5 (Todorov, Erez, and Tassa 2012). In Appendix A, we particularly detail challenging dynamics for UR5 and CartPole, featuring randomized placements of the agent, goal, and obstacles, termed the 'reacher' version. All agents have full access to the current state and are tested using the SAC (Haarnoja et al. 2018) algorithm. We also adjusted the reward structure for both sparse and dense settings, with details in Appendix A.

Environment for generalization. To measure the improvement in generalization, we design a challenging egocentric navigation task using the ViZDoom environment (Kempka et al. 2016), depicted in Figure 3. In the **Seen** environment (Appendix Figure 9-(a)), object locations

Task	Metric	S2D($\mathscr{C}_1$)	S2D($\mathscr{C}_2$)	**S2D($\mathscr{C}_3$)**	Only Sparse	Only dense	D2S($\mathscr{C}_1$)	D2S($\mathscr{C}_2$)	D2S($\mathscr{C}_3$)
Lunar Lander	Perf.	138.71±3.71	63.40±160.55	**168.88**±23.66	142.50±4.25	139.68±14.90	140.75±7.46	130.63±19.69	142.373±15.62
	Sharp.	27.06±36.31	1231.93±2424.61	**7.46**±3.37	8.97±2.83	8.71±4.43	8.95±2.89	8.99±2.97	11.32±3.72
CartPole	Perf.	3.18±4.0	**14.61**±10.96	5.29±7.472	0.14±0.25	3.88±4.63	1.55±0.29	0.38±0.07	0.97±0.19
	Sharp.	0.12±0.24	**0.01**±**0.15**	0.01±0.24	0.08±0.57	0.19±0.03	0.16±0.09	0.05±0.21	0.02±0.17
UR5	Perf.	65.54±10.86	65.69±17.32	**94.15**±4.28	0.00±0.00	64.23±13.03	0.00±0.00	0.00±0.00	0.00±0.00
	Sharp.	0.67±0.01	0.62±0.11	**0.61**±**0.04**	0.09±0.52	0.67±0.01	0.52±0.24	0.56±0.28	0.47±0.20

Table 1: Performance and sharpness metric measured for more than 5 different random seeds in each environment. We highlight the best performance and its sharpness values in bold, confirming that the top-performing **S2D** also resides in the widest minima.

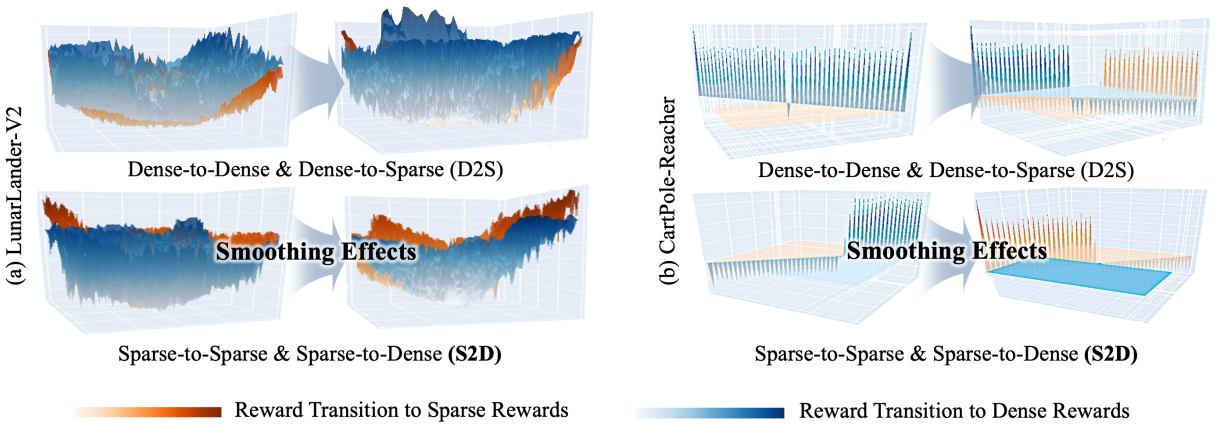


Figure 6: The 3D visualization illustrates the policy loss landscape following a reward transition, which begins with either a sparse or dense reward. It includes two sets of transitions: one transitioning from sparse-to-dense (S2D) and to sparse (Only Sparse), and the other from dense-to-dense (Only Dense) and to sparse (D2S). Notably, the S2D transitions often exhibited more distinct smoothing effects locally compared to others. These effects were noticeable after the transition at T=50 and T=2000 in LunarLander, and at T=3500 in Cartpole. Detailed 3D visualizations for environments are available in Appendix B.

are random, and walls have one of three textures. The **Unseen** environment (Appendix Figure 9-(b)) requires generalization to three new wall textures, distinct from the **Seen** environment. Here, We employed A3C (Mnih et al. 2016).

Results

Performance Enhancement

Sample efficiency and success rate. Experiments are conducted in diverse environments with sparse rewards, and the results are shown in Figure 4, 5 and Table 1. The agents in LunarLander, CartPole-Reacher, ViZDoom-Seen, and ViZDoom-Unseen environments achieve the lowest performance with default sparse rewards. S2D consistently outperforms all other baselines in these settings, achieving better sample efficiency. Even in UR5-Reacher, which is notably more challenging with only sparse rewards than in other environments, the S2D strategy consistently outperforms other baselines. While algorithms based on intrinsic motivation often prioritized exploration over goal attainment in goal-oriented RL tasks, our approach exhibited exceptional results, exhibiting better exploration-exploitation tradeoff while being simple and universally applicable. Importantly, the outcomes associated with D2S consistently fall below those of S2D across all environments, underlining the efficacy of the S2D transition as a curriculum.

Generalization performance. The S2D reward transition outperforms other agents across all dynamic environments requiring generalization as seen in Figure 5-(a), (b). Particularly in ViZDoom-Unseen, where agents encounter drastic visual changes with the emergence of three previously unseen walls, the S2D transition achieves robust generalization and superior performance compared to other baselines.

Post-Transition 3D Policy Loss Landscape

We note a marked smoothing effect at various update points after the reward transition, attributed to the reduction in the depth of local minima, as shown in Table 2 and Figure 6. This effect is predominantly observed under the S2D reward. Such smoothing could aid in overcoming local minima, possibly leading to wide minima. 3D policy loss landscapes of other reward baselines are visualized using the Cross-Density Visualizer in Appendix B. To further verify the smoothing quantitatively, we calculate the depth of local minima of π_θ function, which measures the average of differences between the local maximum and minimum values for a number of updates after the reward transition. The results, as shown in Table 2, demonstrate that the depth primarily decreased for the S2D reward transition model.

Task	Reward Transition	Number of updates after reward transition: phase $\mathcal{P}_1$	$\Delta\mathcal{P}_1$	phase $\mathcal{P}_2$	$\Delta\mathcal{P}_2$	phase $\mathcal{P}_3$	$\Delta\mathcal{P}_3$	phase $\mathcal{P}_4$	$\Delta\mathcal{P}_4$	phase $\mathcal{P}_5$
Lunar Lander	S2D($\mathscr{C}_3$)	0.031±0.01	-0.004	**0.027±0.01**	+0.003	**0.030±0.01**	-0.008	**0.022±0.00**	-0.001	**0.021±0.00**
	D2S($\mathscr{C}_3$)	0.043±0.02	-0.008	0.035±0.00	+0.004	0.039±0.01	+0.004	0.043±0.02	-0.009	0.034±0.01
	only sparse	**0.029±0.01**	0.000	0.029±0.00	+0.005	0.034±0.01	-0.004	0.030±0.00	+0.007	0.037±0.01
	only dense	0.039±0.01	-0.007	0.032±0.01	0.000	0.032±0.01	+0.005	0.037±0.01	-0.002	0.035±0.01
CartPole	S2D($\mathscr{C}_2$)	0.028±0.00	+0.003	**0.031±0.01**	0.000	**0.031±0.00**	+0.005	**0.036±0.00**	-0.013	**0.023±0.00**
	D2S($\mathscr{C}_2$)	0.033±0.01	+0.014	0.047±0.04	+0.001	0.048±0.01	-0.002	0.046±0.02	+0.001	0.047±0.02
	only sparse	**0.027±0.01**	+0.016	0.043±0.01	-0.009	0.034±0.01	+0.007	0.041±0.02	-0.014	0.027±0.00
	only dense	0.033±0.01	+0.021	0.054±0.02	-0.013	0.041±0.02	+0.004	0.045±0.02	-0.001	0.044±0.01
UR5	S2D($\mathscr{C}_3$)	0.084±0.01	+0.016	0.100±0.02	-0.033	0.067±0.01	+0.028	0.095±0.02	-0.062	**0.033±0.01**
	D2S($\mathscr{C}_3$)	0.060±0.03	-0.007	0.053±0.02	-0.005	0.048±0.02	+0.011	0.059±0.02	-0.003	0.056±0.01
	only sparse	**0.059±0.02**	-0.013	**0.046±0.00**	0.000	**0.046±0.01**	+0.031	0.077±0.00	-0.031	0.046±0.01
	only dense	0.079±0.02	-0.012	0.067±0.02	-0.009	0.058±0.02	-0.005	**0.053±0.01**	+0.005	0.058±0.01

Table 2: Comparison of the depth of local minima in policy loss following reward transitions. A lower depth of local minima indicates a smoother terrain. Phase $\mathcal{P}_i$ indicates the number of updates, which are $(T = 50, 400, 800, 1200, 1600)$ for LunarLander, and $(T = 50, 1000, 2000, 3000, 4000)$ for others.

Results of Wide Minima

We assess the end-of-training convergence of the neural networks guided by S2D using sharpness metrics and contrast this with the baselines. Areas of lower sharpness correspond to wide minima, which can enhance generalization performance. As demonstrated in Table 1, only the agents guided by S2D reward transition that converge to the widest minima exhibit superior performance in challenging environments.

Discussion and Analyses

In the following analyses, we clearly address the three pivotal questions raised in our experimental framework:

Performance enhancement. As shown in Figure 4, Figure 5, and Table 1, S2D surpassed other reward baselines. In our experiments comparing intrinsic motivation methods, we observed that agents driven by such algorithms tend to prioritize state coverage over achieving specific goals in goal-oriented RL tasks. This suggests that while they may excel in exploration, the S2D transition approach more effectively balances exploration and exploitation, thus ensuring proper goal acquisition. Additionally, we explored the optimal timing for reward transition through ablation studies. While this timing is unique for each environment, it was, in all cases, near a quarter of the total training time. Particularly challenging tasks like UR5-Reacher required longer periods of free exploration compared to relatively simpler ones, such as LunarLander. This echoes the critical early learning phases observed in infants.

Post-transition 3D policy loss landscape. Our 3D visualizations reveal that S2D predominantly smooths the landscape. Although our main experiments are based on SAC, we also tested other algorithms such as PPO (Schulman et al. 2017) and DQN (Mnih et al. 2013) for more comprehensive analyses. This smoothing effect was also uniquely seen with the S2D reward transition in additional gridworld experiments, as discussed in Appendix C.

In UR5, minima depth for S2D decreased significantly from 0.095 to 0.033 in phase $\mathcal{P}_5$. This may relate to its higher performance than only dense rewards particularly at a later stage, as in Figure 4-(b). Conversely, in CartPole and LunarLander, performance quickly improved after the reward transition, reflecting S2D's overall low local minima.

Correlation between wide minima and toddler-inspired reward transition. In Table 1, S2D-guided agents converged to the widest minima with the highest performance in the LunarLander and CartPole-Reacher, where agents receiving only sparse rewards achieved nonzero performance. Conversely, in the UR5-Reacher, agents with only sparse rewards showed zero performance. This implies that in sparse reward situations, premature convergence into wide minima can lead to gradient stagnation and retain low sharpness, with high variance. Yet, S2D, compared to only dense rewards, still posts the highest performance with the lowest sharpness, suggesting its alignment within wide minima.

Conclusion

Inspired by toddler developmental learning, our research pioneers a shift from static, single-density to dynamic reward transitions in goal-oriented RL. This toddler-inspired approach demonstrates notable effects of transitions on learning dynamics in RL. We examine its implications across various scenarios, focusing on its efficiency. Using the Cross-Density Visualizer, we observe the primary smoothing effects on the policy loss landscape during the S2D transition. Sharpness metrics further confirm the smoothing effects of S2D, guiding agents towards wider minima to improve generalization. This blend of biological and artificial paradigms may lead to robust, high-performance learning systems.

Limitations. While our study focuses on understanding the implications of S2D reward transition, we haven't provided an automatic method for finding an optimal transition yet. Nonetheless, our preliminary research on criteria sets the stage for future development of automated methods.

Acknowledgments

The authors would like to express their sincere gratitude to Inwoo Hwang, Changhoon Jeong, Moonhoen Lee, and Dong-Sig Han for their insightful discussions and valuable suggestions on the early drafts of this paper. This work was partly supported by the IITP (2021-0-02068-AIHub/15%, 2021-0-01343-GSAI/10%, 2022-0-00951-LBA/15%, 2022-0-00953-PICA/25%) and NRF (RS-2023-00274280/10%, 2021R1A2C1010970/25%) grant funded by the Korean government.

References

Achille, A.; Rovere, M.; and Soatto, S. 2018. Critical learning periods in deep networks. In *International Conference on Learning Representations*.

Andrychowicz, O. M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1): 3–20.

Aubret, A.; Matignon, L.; and Hassas, S. 2019. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*.

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; et al. 2020. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*.

Bambach, S.; Crandall, D.; Smith, L.; and Yu, C. 2018. Toddler-inspired visual object learning. *Advances in neural information processing systems*, 31.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML '09*.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

De Kleijn, R.; Sen, D.; and Kachergis, G. 2022. A Critical Period for Robust Curriculum-Based Deep Reinforcement Learning of Sequential Action in a Robot Arm. *Topics in Cognitive Science*, 2(2): 311–326.

Dong, Q.; Gong, S.; and Zhu, X. 2017. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 1851–1860.

Du, B.; Gao, X.; Hu, W.; and Li, X. 2021. Self-Contrastive Learning with Hard Negative Sampling for Self-Supervised Point Cloud Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 3133–3142. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Florensa, C.; Held, D.; Geng, X.; and Abbeel, P. 2018. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, 1515–1528. PMLR.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Gibson, E. J. 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 39(1): 1–42.

Goodfellow, I. J.; Vinyals, O.; and Saxe, A. M. 2014. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*.

Gopnik, A.; Meltzoff, A. N.; and Kuhl, P. K. 1999. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co.

Gopnik, A.; O'Grady, S.; Lucas, C. G.; Griffiths, T. L.; Wente, A.; Bridgers, S.; Aboody, R.; Fung, H.; and Dahl, R. E. 2017. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, 114(30): 7892–7899.

Graves, A.; Bellemare, M. G.; Menick, J.; Munos, R.; and Kavukcuoglu, K. 2017. Automated curriculum learning for neural networks. In *international conference on machine learning*, 1311–1320. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hacohen, G.; and Weinshall, D. 2019. On The Power of Curriculum Learning in Training Deep Networks. *ArXiv*, 2.

Jastrzębski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2018. Finding Flatter Minima with SGD.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard Negative Mixing for Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21798–21809. Curran Associates, Inc.

Kempka, M.; Wydmuch, M.; Runc, G.; Toczek, J.; and Jaśkowski, W. 2016. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE conference on computational intelligence and games (CIG)*, 1–8. IEEE.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*.

Knox, W. B.; Allievi, A.; Banzhaf, H.; Schmitt, F.; and Stone, P. 2023. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316: 103829.

Ladosz, P.; Weng, L.; Kim, M.; and Oh, H. 2022. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85: 1–22.

Laud, A. D. 2004. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

MacKay, D. J. 1992. Information-based objective functions for active data selection. *Neural computation*, 4(4): 590–604.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Narvekar, S.; and Stone, P. 2020. Generalizing curricula for reinforcement learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, 278–287. Citeseer.

Oudeyer, P.-Y.; and Smith, L. B. 2016. How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, 8(2): 492–502.

Park, J.; Park, K.; Oh, H.; Lee, G.; Lee, M.; Lee, Y.; and Zhang, B.-T. 2021. Toddler-Guidance Learning: Impacts of Critical Period on Multimodal AI Agents. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 212–220.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Piaget, J.; Cook, M.; et al. 1952. *The origins of intelligence in children*, volume 8. International Universities Press New York.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shonkoff, J.; and Phillips, D. 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. eric. ed. gov. *National Academy of Sciences Press: Washington DC. Accessed on May*, 8: 2015.

Taylor, M. E.; and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7).

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 5026–5033. IEEE.

Turchetta, M.; Kolobov, A.; Shah, S.; Krause, A.; and Agarwal, A. 2020. Safe reinforcement learning via curriculum induction. *Advances in Neural Information Processing Systems*, 33: 12151–12162.

Weinshall, D.; Cohen, G.; and Amir, D. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, 5238–5246. PMLR.

Zhang, B.-T. 1994. Selecting a Critical Subset of Given Examples during Learning. In *International Conference on Artificial Neural Networks*, 517–520. Springer.