Learning to See Through a Baby's Eyes: Early Visual Diets Enable Robust Visual Intelligence in Humans and Machines

Yusen Cai[1] Bhargava Satya Nunna[1,2] Qing Lin[1,*] Mengmi Zhang[1,*]

[1]Nanyang Technological University, Singapore
[2]Indian Institute Of Technology Madras
[*]Co-corresponding authors
Address correspondence to mengmi.zhang@ntu.edu.sg

Abstract

Newborns perceive the world with low-acuity, color-degraded, and temporally continuous vision, which gradually sharpens as infants develop. To explore the ecological advantages of such staged "visual diets", we train self-supervised learning (SSL) models on object-centric videos under constraints that simulate infant vision: grayscale-to-color (C), blur-to-sharp (A), and preserved temporal continuity (T)—collectively termed CATDiet. For evaluation, we establish a comprehensive benchmark across ten datasets, covering clean and corrupted image recognition, texture–shape cue conflict tests, silhouette recognition, depth-order classification, and the visual cliff paradigm. All CATDiet variants demonstrate enhanced robustness in object recognition, despite being trained solely on object-centric videos. Remarkably, models also exhibit biologically aligned developmental patterns, including neural plasticity changes mirroring synaptic density in macaque V1 and behaviors resembling infants' visual cliff responses. Building on these insights, CombDiet initializes SSL with CATDiet before standard training while preserving temporal continuity. Trained on object-centric or head-mounted infant videos, CombDiet outperforms standard SSL on both in-domain and out-of-domain object recognition and depth perception. Together, these results suggest that the developmental progression of early infant visual experience offers a powerful reverse-engineering framework for understanding the emergence of robust visual intelligence in machines. All code, data, and models will be publicly released.

1. Introduction

Newborns perceive a world that is blurry, desaturated, and continuously unfolding in time [68, 115, 117, 118, 126, 127]. Over the first year of life, visual acuity sharpens and color sensitivity emerges as their visual inputs mature [5, 25, 70, 78, 103, 110]. This early limitation is not a defect but a developmental scaffold, enabling the brain to organize sensory experience into stable and generalizable representations [8, 12, 27, 30, 104–106, 116, 138]. When this developmental process is disrupted, perceptual deficits can arise. For instance, the absence of early low-acuity input has been linked to lasting deficits in face processing [36, 69, 86, 115], and immature photoreceptors bias infants toward luminance-based rather than chromatic cues [117].

In contrast, artificial visual systems in AI are typically trained on fully detailed, static images. To increase data diversity and improve model robustness, random data augmentations are commonly applied [17, 26, 43, 46, 89, 137, 142, 146]. While such training regimes achieve impressive performance on standard benchmarks, they remain ecologically invalid, overlooking how perception in nature develops through structured and temporally coherent visual experience [67, 102, 109, 129]. As a result, modern vision models struggle to generalize to corrupted or occluded images [23, 29, 45, 72, 101, 113, 123, 125]. This limitation constrains their reliability in high-stakes real-world applications, including autonomous driving, robotics, and embodied human–AI interaction [77, 120, 131, 132, 134, 141, 145].

What enables human perceptual robustness, and how might we reverse-engineer it? Decades of developmental research suggest two key factors. First, the structure of visual experience — including spatial statistics, infant-perspective object distributions, and temporal continuity — guides learning toward generalizable shape-based representations rather than brittle cue-specific solutions [3, 24, 34, 51, 74, 76, 85, 96, 103, 104, 127, 138]. Second, an initially degraded (blurred, low-chroma) input may serve as an adaptive scaffold, biasing developmental

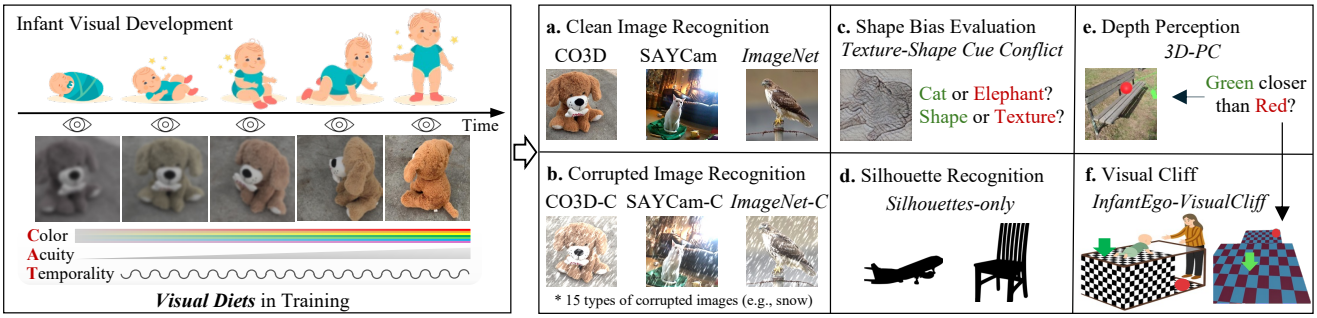

Figure 1. **Illustration of developmental visual diets and overview of evaluation benchmarks.** The **left** panel depicts stages of infant development over time along with corresponding characteristics of visual perception. The 3 axes below represent the key regularities of infant visual development that underpin our work: Color, Acuity, and Temporality. The *Color* diet (CDiet) models the progression from color-degraded to richly chromatic scenes as color vision matures. The *Acuity* diet (ADiet) reflects the transition from blurry to sharp perception as visual resolution improves. The *Temporality* diet (TDiet) captures infants' exposure to smoothly evolving visual scenes over short time windows. In the **right** panel, regular font indicates in-domain tasks, while *italic font* denotes out-of-domain tasks. Panels (a–d) assess object recognition: (a) clean image recognition on CO3D [90], SAYCam [82], and ImageNet [92]; (b) corrupted image recognition across 15 corruption types following ImageNet-C [45]; (c) shape-bias evaluation using the Texture–Shape Cue Conflict dataset [35], testing whether classification aligns with shapes (green) or textures (red); and (d) silhouette recognition using the Silhouettes-Only dataset [35]. Panels (e–f) evaluate depth perception: (e) judging whether the green arrow is closer than the red ball in the 3D-PC dataset [63]; and (f) predicting which side (green arrow or red ball) appears closer from the infant egocentric perspective in the Visual Cliff paradigm [37].

trajectories toward useful inductive bias with lasting effects on perceptual organization [71, 115–118].

Prior works in machine vision have leveraged some of these principles from developmental psychology. For example, some studies train models on longitudinal egocentric videos collected from head-mounted cameras worn by infants [81, 82, 99], but these works do not account for key characteristics of infant visual systems, such as low acuity and limited color sensitivity. More recent works have explored individual visual diets, such as progressive blur or color exposure [115, 117, 118], yet these models rely on fully supervised training with thousands of labeled examples, which is not ecologically plausible. To address these gaps, as shown in **Fig. 1**, we design developmentally inspired visual diets that capture key features of infant vision, following a progression from grayscale to color (**CDiet**), blur to sharp (**ADiet**), and preserving temporal continuity (**TDiet**), collectively termed **CATDiet**. We train self-supervised learning (SSL) models on object-centric videos under these constraints and examine how each diet and their combination shapes the emergence of robust visual representations.

In parallel, SSL has recently emerged as a powerful approach for learning visual representations without labels. Prominent SSL methods [10, 14, 20, 32, 44] include SimCLR [18], MoCo [42], BYOL [38], MAE [43], and DINO [15]. However, these approaches largely rely on random data augmentations, leaving the potential of developmentally inspired visual diets underexplored. To investigate this, we introduce a comprehensive benchmark spanning 10 datasets for object recognition, texture–shape conflict tests, silhouette recognition, and depth perception, summarized in **Fig. 1**. On these benchmarks, CATDiet models, as well as their individual components, not only enhance robustness to visual corruptions but also exhibit developmental signatures consistent with biological vision, including changes in neural plasticity, the emergence of depth sensitivity, and behavioral patterns resembling infants in the visual cliff paradigm [37].

To further leverage CATDiet, we introduce CombDiet, which starts training with CATDiet before transitioning to standard SSL while maintaining temporal continuity. This hybrid approach combines the ecological grounding of early visual experience with the efficiency of mature vision, consistently outperforming standard SSL on both in-domain and out-of-domain image recognition and depth perception tasks. Our main contributions are as follows:

1. We introduce CATDiet, a developmentally inspired visual diet in SSL that emulates the progression of infant vision through 3 staged constraints: grayscale-to-color (C), blur-to-sharp (A), and preserved temporal continuity (T).

2. We establish comprehensive benchmarks across 10 datasets covering object recognition and depth perception. The evaluation suite includes corrupted image recognition, texture–shape cue conflict tests, silhouette recognition, depth-order classification, and the visual cliff paradigm. This framework allows systematic, quantitative assessment of how individual visual diets and their combinations improve visual robustness.

3. Remarkably, even when trained exclusively on object-centric videos—without supervision from infant behavior or monkey neural data—CATDiet exhibits

developmental signatures aligned with biological vision, including neural plasticity changes consistent with synaptic density in macaque V1, early emergent depth sensitivity, and behavioral patterns resembling infants' responses in the visual cliff paradigm.

4. Building on these insights, we introduce CombDiet, which initializes standard SSL with CATDiet while maintaining temporal continuity. CombDiet substantially outperforms standard SSL across our benchmarks, underscoring the practical benefits of early visual diets for real-world computer vision applications.

2. Related Works

Developing AI models inspired by developmental psychology. A growing body of research draws inspiration from developmental psychology to inform AI model design. Existing approaches can be broadly categorized into two directions: (1) pretraining on egocentric videos of children to capture spatial statistics, multimodal associations, and infant-perspective object distributions [81, 82, 99, 107], and (2) modeling specific developmental transitions—such as from blur to sharp or grayscale to color vision—under controlled, supervised settings [7, 115–118]. While these studies provide valuable insights, they have key limitations. Egocentric videos approximate viewing statistics but overlook crucial properties of infants' visual systems, such as low spatial acuity and limited color sensitivity, which we explicitly model. Moreover, training directly on raw video streams makes it difficult to disentangle which factors contribute to generalizable object representations; our benchmark enables systematic, quantitative evaluation of individual and combined visual diets. Prior evaluations also emphasize clean-image recognition, neglecting model robustness and generalization, which we address through extensive out-of-domain testing. Finally, most prior works rely on supervised learning, reducing ecological validity and scalability; in contrast, our framework embeds developmental principles within a self-supervised paradigm to enable fully label-free, ecologically grounded learning.

Self-Supervised Learning. Self-Supervised Learning (SSL) has become a dominant paradigm for learning visual representations without manual annotations. Existing SSL approaches can be broadly grouped into four categories: (1) contrastive methods [19, 21, 22, 44, 130, 135], such as SimCLR [18] and MoCo [42], which pull together augmented views of the same image while pushing apart different ones; (2) non-contrastive methods [10, 32, 56], including BYOL [38], SimSiam [20], and Barlow Twins [139], which achieve invariance without negative pairs through architectural or loss-based constraints that prevent representational collapse; (3) generative methods [111, 122], such as autoencoders [48] and masked autoencoders [43], which reconstruct missing or corrupted inputs; and (4) clustering-based methods [54, 80, 100], such as SwAV [14] and DINO [15], which align features with evolving cluster prototypes. While these methods rely on random data augmentations to enrich training signals, they overlook the developmental structure of visual experience. Building on video-based SSL frameworks that leverage temporal continuity across frames [6, 33, 40, 59, 79, 83, 87, 95, 112, 119, 128], we introduce TDiet, which explicitly enforces temporal alignment between adjacent frames to capture the natural continuity of visual experience. Unlike prior video-based SSL models, which are typically evaluated only on video understanding benchmarks, we also assess them on corrupted image recognition and depth perception tasks. Furthermore, by integrating TDiet with the CDiet and ADiet, our combined CATDiet framework achieves a synergistic performance improvement that surpasses the contribution of any individual component.

Improving Model Robustness to Visual Corruptions. Out-of-distribution (OOD) generalization in visual representation learning remains a longstanding challenge [9, 11, 16, 52, 64, 66, 93, 140, 143, 144]. Existing approaches have explored a wide range of strategies, including specialized training objectives [2, 49, 91, 94], large-scale and diverse datasets [80, 100], advanced data augmentations [46, 47, 137, 142], domain-invariant feature learning [61, 62, 73, 75, 98, 124], generative modeling [50, 121], and architectural innovations [4, 13, 53, 58, 97, 108, 114]. While these approaches improve robustness on benchmarks such as ImageNet-C [45], they often depend on extensive artificial data augmentations or non-ecological training objectives, rather than fostering emergent generalization from limited and naturalistic visual experience—as seen in humans [72]. In contrast, we examine how early staged visual diets, inspired by human visual development and learned through self-supervision, can enhance model robustness across diverse in-domain and out-of-domain benchmarks spanning 10 datasets.

3. Our Proposed Developmental Visual Diets

We propose CombDiet (**Fig. 2d**), a two-phase self-supervised learning (SSL) framework that embeds principles of human visual development into both the data curriculum and the learning objectives. In the first phase, CATDiet serves as a warm-up stage that mirrors infants' first-year visual progression, characterized by rapid perceptual development toward near-mature vision. This phase spans the initial 30% of training epochs and integrates two data curricula—Color-Diet (CDiet) and Acuity-Diet (ADiet)—with a temporal regularization objective, Temporality-Diet (TDiet), to emulate the temporal continuity inherent in early visual experience. In the second phase, Standard Diet (SDiet) represents mature

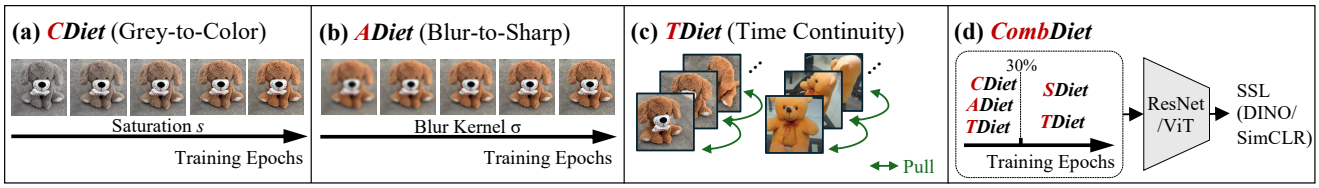

Figure 2. **Overview of our proposed developmental visual diets.** *CAT*Diet (panels a–c) integrates 3 individual visual diets detailed in **Sec. 3.1**: (a) CDiet, in which image saturation gradually increases as chromatic information is progressively introduced throughout training; (b) ADiet, where the standard deviation σ of Gaussian blur kernels decreases, enhancing spatial details over time; and (c) TDiet, which encourages representations of adjacent views of the same object to be closer, capturing temporal continuity in object-centric videos. (d) CombDiet extends CATDiet to a more general setting. In the first phase, CATDiet serves as a warm-up stage spanning the initial 30% of training epochs. In the second phase, CombDiet transitions to the Standard Diet (SDiet) while retaining Temporality-Diet (TDiet). SDiet corresponds to the standard data augmentation pipeline used in conventional SSL training regimes. We evaluate CombDiet using 2 representative SSL methods (SimCLR [18] and DINO [15]) with 2 widely adopted backbones: ResNet [41] and ViT [31].

visual experience. CombDiet transitions to SDiet while retaining TDiet, thereby maintaining temporal coherence throughout learning.

3.1. Phase 1 - CATDiet

Color-Diet (CDiet). To mimic the grey-to-color progression, we design a five-stage saturation schedule (**Fig. 2a**), where each stage specifies a blend ratio s between the fully colored image I_c and its grayscale counterpart I_g. From the first to fifth stages, s is sampled within (0.20, 0.36), (0.36, 0.52), (0.52, 0.68), (0.68, 0.84), and (0.84, 1.0), respectively. For example, in the first stage, for any given training image, a random s is sampled within (0.2, 0.36) and used to blend its grayscale and colored versions: $sI_c + (1-s)I_g$. Stage durations $[10, 7, 6, 5, 2]$, measured in training epochs, decrease across stages. This design choice reflects the rapid increase in chromatic sensitivity during early infancy [25, 28, 68, 103, 117, 126].

Acuity-Diet (ADiet). For the blur-to-sharp curriculum, we define a five-stage Gaussian blur schedule (**Fig. 2b**). Across stages one to five, each training image is blurred with standard deviations $\sigma \in [4, 3, 2, 1, 0]$, corresponding to kernel sizes $[25, 19, 13, 7, 1]$ pixels for images of size 224×224. Stage durations $[10, 6, 6, 3, 5]$, measured in training epochs, decrease across stages. This design reflects developmental psychology findings that visual acuity increases approximately exponentially during the first year of life [78, 110]. In CATDiet, the ADiet and CDiet are integrated by interleaving the Gaussian blur schedule of ADiet with the saturation schedule of CDiet, with each schedule maintaining its own stage durations.

Temporality-Diet (TDiet). We introduce a temporal alignment objective (**Fig. 2c**) to capture temporal continuity. Adjacent video frames are closely related both temporally and spatially, often depicting the same object undergoing small, continuous changes in viewpoint. This temporal coherence provides an intrinsic, free supervisory signal, encouraging SSL models to learn view-invariant representations by pulling the embeddings of adjacent frames closer [127]. Combined with CDiet and ADiet, which apply augmentations to each frame, the augmented versions of adjacent frames are grouped as a positive set. The training objective is therefore to bring together the representations of adjacent frames and their augmented variants.

3.2. Phase 2 - SDiet and TDiet

CombDiet allocates the first 30% of training epochs to Phase 1 and the remaining 70% to Phase 2, with these hyperparameters determined via grid search to optimize performance. In Phase 1, CATDiet emulates the progression of early visual diets during the first year of life, while Phase 2 transitions to SDiet to approximate adult visual experience. In **SDiet**, standard data augmentation techniques from the 2 representative SSL methods described below are applied to training images. The TDiet objective remains unchanged in Phase 2, consistently encouraging alignment of representations between adjacent frames and their augmented versions throughout training.

SSL Methods and Model Backbones. Ideally, our visual diets could be applied to all SSL methods and model backbones, but exhaustively evaluating all combinations is computationally prohibitive. We therefore select 2 representative SSL methods, SimCLR [18] and DINO [15]. Unlike generative SSL methods such as MAE [39], which reconstruct missing or corrupted pixels and are less ecologically valid, SimCLR and DINO are widely used and yield more biologically plausible SSL representations [39, 55, 84, 133, 136, 147]. Given their distinct training objectives, we adapt TDiet to each method. In SimCLR, TDiet pulls adjacent frames and their augmented views closer in embedding space while keeping the original negative pairs to push apart representations of non-adjacent frames within the same video or frames from different videos. In DINO, TDiet aligns features of adjacent

frames and their augmented views through evolving cluster prototypes from the student and teacher networks. For each SSL method, we experiment with 2 common backbones: the 2D-CNN ResNet [41] and the transformer-based ViT [31], yielding 4 model variants denoted as [SSL method]-[backbone] (e.g., SimCLR-ResNet).

Implementation Details. All models are trained with a batch size of 64 and an image resolution of 224×224. We use the AdamW optimizer [65] with a learning rate of 5×10^{-4}, weight decay of 1×10^{-4}, and a cosine annealing schedule with a 10-epoch warm-up. Experiments are conducted on NVIDIA RTX A6000 and RTX 6000 Ada Generation GPUs. Additional implementation details, and parameter analyses are provided in **Sec. S1**.

4. Benchmarking Developmental Visual Diets

In this section, we present a comprehensive benchmark for evaluating developmental visual diets. It encompasses 10 datasets, multiple evaluation metrics across diverse downstream tasks, and baseline models for comparison with our proposed developmental visual diets.

4.1. Datasets

CO3D [90]: The CO3D dataset comprises real-world, object-centric videos capturing full 360° azimuthal rotations around individual objects. We select 10 object categories with about 4,500 instances in total. Training and test splits are defined at the object-instance level within each category. During training, we uniformly sample 10 frames per video to span the complete viewing circle, corresponding to roughly 36° between adjacent views. For testing, we randomly sample 2 frames per video to ensure viewpoint diversity while minimizing redundancy.

CO3D-C: To evaluate model robustness in object recognition, we construct a corrupted variant of the CO3D test set, using the distortion operators defined in the ImageNet-C protocol [45]. CO3D-C comprises 15 corruption types across 4 families—noise, blur, weather, and digital artifacts—each applied at 5 severity levels.

3D-PC [63]: The 3D-PC Depth Order dataset contains 4,500 images of rendered 3D scenes. Each image depicts a green arrow (camera viewpoint) and a red ball. The task is formulated as a binary classification problem, where the model predicts whether the arrow is closer than the ball (**Fig. 1e**). We adopt the official training and test splits.

IEVC: The classical visual cliff paradigm in developmental psychology places infants on a glass platform with one side appearing shallow (texture directly beneath the glass) and the other deep (texture several centimeters below). Infants' hesitation to cross the "cliff" marks the emergence of depth perception. To simulate this, we render 3 infant-perspective views in Blender facing the deep side (**Fig. 1f** and **Fig. 4c**). Similar to 3D-PC, we formulate this task as a binary depth-order classification task for the models.

SAYCam (SAY) [107]: SAY is a longitudinal egocentric video dataset capturing the daily visual experiences of 3 infants (S, A, and Y) from 6 to 32 months of age, providing spatial statistics, infant-perspective object distributions, and temporal continuity reflective of early visual development. In our experiments, we use recordings from child S, which include an annotated subset of video frames across 26 object categories. The unannotated portion is used for self-supervised pretraining, while the annotated frames serve as the linear probing dataset. From each video, we sample a 2-second clip every 2 minutes to minimize redundancy, with each clip sampled at 5 fps, yielding approximately 35,000 frames for pretraining.

SAYCam-C (SAY-C): A corrupted extension of SAY is constructed using the same 4 corruption families as the CO3D-C dataset. This dataset evaluates model robustness in object recognition under corrupted egocentric views.

ImageNet-21K (IN) [92]: IN is a large-scale naturalistic image dataset containing over 14 million images across 21,000 object categories. For our experiments, we select 15 overlapping classes between IN and SAY. The list of overlapping classes is provided in **Sec. S1**. **ImageNet-C (IN-C)**: Corrupted IN is constructed using the same 4 corruption families as CO3D-C.

Texture–Shape Cue Conflict (TSCC) [35]: TSCC is a diagnostic dataset where each image fuses the global shape of one object category with the local texture of another via style transfer (**Fig. 1c**). It evaluates whether SSL models rely primarily on shape or texture for object recognition. We select 3 overlapping object classes between TSCC and SAY for evaluation.

Silhouettes-Only (Sil-O) [35]: Sil-O is a texture-free variant of IN where each object is rendered as a black silhouette on a white background (**Fig. 1d**). We select the same 3 overlapping classes as in TSCC. This dataset assesses shape-based recognition independent of color.

4.2. Baselines

Baselines for CDiet and ADiet. We define 4 baselines by altering the order of their staged curricula. For brevity, we denote each baseline together with its corresponding visual diet as **[Diet]-[Baseline]**; for example, **C-REV** indicates the REV baseline applied to CDiet. **Reverse Order (REV).** The progression direction in the staged schedule is reversed; for instance, instead of progressing from grayscale to color in CDiet, **C-REV** applies the five-stage saturation schedule from color to grayscale. **Shuffled Order (SHF).** The original stage ordering is discarded, and data from all 5 stages are merged into a single pool, with each mini-batch randomly sampled from this pool during SSL pretraining. **First-Only (FO).** To limit SSL models to early-stage

inputs, they are pretrained only on the first stage of each visual diet; for example, in **C-FO**, only the lowest blend ratio s from stage 1 is applied across all 5 stages in CDiet. **Last-Only (LO).** LO uses all the original images without any color or blur modifications. For example, **C-LO** uses the original full-color images for pretraining.

Baselines for TDiet. We include a single baseline, **Non-Smooth**, in which the objective pulls together only cropped views from the same frame, without encouraging similarity between neighboring frames.

Baselines for CATDiet. We adopt the same 4 baselines as for CDiet and ADiet, applying each baseline to its respective schedule individually and then combining them to form the corresponding baselines, such as **CAT-REV**.

Baselines for CombDiet. We compare **CombDiet** against two controlled baselines: **Shuffled Order (SHF).** This variant follows the same two-phase schedule as CombDiet, with CAT-SHF used in Phase 1 followed by SDiet in Phase 2. It isolates the effect of the developmental order while maintaining the same data distribution, architecture, and training duration as CombDiet. **Standard SSL (STD).** We include standard SSL baselines, SimCLR and DINO. For fair comparisons, we use the same backbones as in CombDiet but retain their conventional training paradigm with standard data augmentations and loss objectives.

4.3. Evaluation Metrics

Metrics for object recognition. **Top-1 accuracy (Acc)** denotes top-1 classification accuracy. The **mean Corruption Error (mCE)** [45] is computed as the average normalized error across all corruption types and severity levels, with lower values indicating stronger model robustness. See **Sec. S1** for details. **Shape-Bias (S-Bias)** quantifies perceptual alignment with shape cues in the TSCC dataset. It is defined as the proportion of shape-consistent predictions among all images classified as either shape- or texture-consistent by any of the three models (CombDiet, SHF, STD). Higher S-Bias indicates stronger reliance on global shape cues over local textures.

Metrics for quantifying network connectivity. **Fisher Information Matrix (FIM)** is defined as the trace of the Fisher Information Matrix [1], computed from the gradients of the network parameters. Intuitively, it reflects the network's sensitivity to small parameter perturbations and its adaptability during SSL pretraining.

Metrics for depth-order classification. **Depth Accuracy (dAcc)** denotes the binary classification accuracy on the depth-order prediction task. dAcc of 0.5 corresponds to chance performance, reflecting random guesses on whether the green arrow is closer than the red ball.

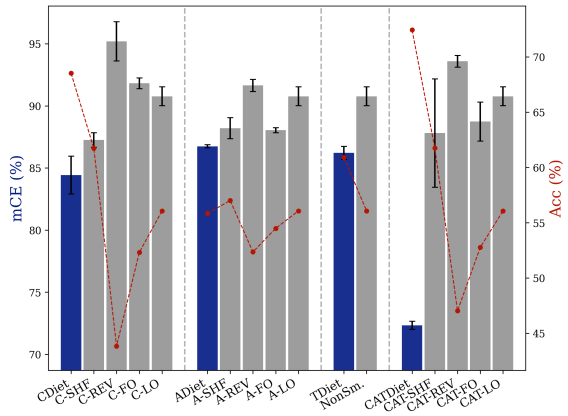

Figure 3. **Object recognition performance on CO3D (clean) and CO3D-C (corrupted) datasets for SimCLR-ResNet pretrained on CATDiet and its individual diets.** Bars show mCE (↓, left axis), where blue and gray bars denote our proposed visual diets and their corresponding baselines; the dashed red line indicates Acc (↑, right axis). Error bars represent the standard error of mCE over three runs. The four panels correspond to different attributes of the proposed visual diets: Color, Acuity, Temporality, and their combination (see **Sec. 5.1**).

5. Results

5.1. Evaluation of CATDiet on Object Recognition

We pretrain 4 SSL models on the integrated CATDiet as well as its individual diets using the CO3D training set, followed by training 10-way linear probes on the same set. The models and their baselines are then evaluated on clean CO3D test images and corrupted CO3D-C images. Results for SimCLR-ResNet are shown in the main text, with other models presented in **Sec. S2**. Consistent observations and analyses apply across all models.

Natural progression order matters and incorrect progression order impairs model robustness. From the first two panels of **Fig. 3**, both CDiet and ADiet models pretrained with the SHF or REV exhibit higher mCE compared to their original staged curricula. For example, CDiet outperforms the best-performing baseline **C-SHF** by 2.9% in mCE. This indicates that maintaining the original schedule in visual diets is critical for learning robust object representations, as it provides the correct inductive bias. Interestingly, models pretrained with REV even underperform SHF (e.g., **C-REV** with mCE of 95.2% versus **C-SHF** with mCE of 87.3%), suggesting that incorrect progression orders strongly disrupt the model's ability to capture meaningful visual features.

Early limitations in visual diets are beneficial for developmental scaffolding but insufficient on their own. To probe the effect of early-stage and late-stage diets on model robustness, we compare the FO and LO baselines against our proposed CDiet and ADiet (first two panels of **Fig. 3**). Pretraining on LO is inferior to the full staged

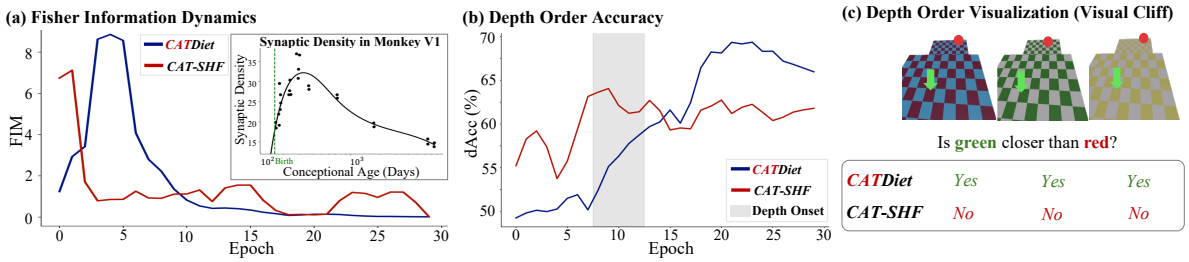

Figure 4. **Signature developmental patterns observed in SimCLR-ResNet pretrained on CATDiet from CO3D (a), 3D-PC (b), and IEVC (c).** (a) The trace of the Fisher Information Matrix (FIM) for the SSL model gradients [1] is plotted across pretraining epochs, capturing the sensitivity of network outputs to small weight perturbations. The inset shows synaptic density changes in macaque primary visual cortex (V1) [88], highlighting a similar rise-and-fall pattern for the model pretrained on CATDiet (blue) compared to FIM changes in CAT-SHF (red). (b) Binary classification accuracy on a depth-order task as a function of pretraining epochs. Blue and red curves correspond to CATDiet and SHF, respectively. The shaded region marks the period of rapid accuracy increase in dAcc for CATDiet. (c) Simulated Visual Cliff experiment. The top row shows egocentric views from an infant's perspective crawling on a glass platform (see **Sec. 4.1**). The table below summarizes model responses for CATDiet and CAT-SHF to the binary question "Is the green arrow closer than the red ball?" ("yes" indicates that the green arrow is closer).

curriculum, suggesting that limited early-stage exposure helps scaffold representation learning. For example, CDiet achieves an mCE of 84.4%, compared to 90.8% for C-LO. However, FO alone does not match the performance of our proposed diet (e.g., CDiet vs. C-FO and ADiet vs. A-FO), demonstrating that early constraints facilitate learning but require subsequent stages to fully develop robust representations. These empirical results echo two observations from developmental psychology. First, children who begin visual experience with relatively high acuity due to early cataract removal can discriminate faces based on local features but fail to detect their configural changes [69, 115]. Second, children whose vision begins with mature cone cells due to early cataract removal show a marked decrease in recognizing grayscale images relative to color images [117].

Models achieve strong accuracy on clean images while exhibiting robustness to image corruptions. From the first two panels of **Fig. 3**, across CDiet and ADiet, models pretrained with our proposed diets not only perform competitively well with or better than all baselines (REV, SHF, FO, LO) in terms of Acc on clean images, but also significantly outperform all baselines in terms of mCE on corrupted images. This demonstrates that developmental-inspired training simultaneously enhances object recognition accuracy and model robustness.

Temporal continuity provides free and useful regularization to learn robust object representations. In TDiet (third panel of **Fig. 3**), including the training objective of temporal continuity leads to consistent gains in both Acc and mCE compared to the Non-Smooth baseline. By encouraging similarity between neighboring frames, TDiet effectively regularizes the learned representations, improving robustness to natural image corruptions.

Integrated CATDiet achieves more than its individual components. In the fourth panel of **Fig. 3**, CATDiet achieves a substantially lower mCE of 72.3% and a higher Acc of 72.4% compared to its individual components (CDiet: 84.4% in mCE and 68.5% in Acc, ADiet: 86.7% in mCE and 55.8% in Acc, TDiet: 86.2% in mCE and 60.9% in Acc). This demonstrates that the combined curriculum provides synergistic benefits beyond the sum of its parts. Furthermore, we compare CATDiet against its baseline variants in both mCE and Acc. Consistent with observations in CDiet and ADiet individually, the baselines underperform CATDiet, with performance gaps even larger than those observed for the individual diets, further highlighting the importance of integrating developmentally inspired diets.

Changes in network connectivity of SSL models pretrained on CATDiet mirror synaptic density changes in macaque V1 over development. As shown in **Fig. 4a**, network connectivity measured via FIM for CATDiet closely follows synaptic density changes in macaque primary visual cortex across the lifespan [88]. FIM rises sharply during early pretraining, peaks around the 5th epoch, and gradually declines, indicating an initial phase of high plasticity followed by progressive weight consolidation. This suggests that CATDiet undergoes an early exploratory stage of representation formation before stabilizing into robust feature embeddings. In contrast, CAT-SHF shows a monotonically decreasing curve, implying premature convergence. These results indicate that a structured developmental diet balances exploration and exploitation, with the model first exploring the embedding space and later consolidating robust representations through weight pruning.

5.2. Evaluation of CATDiet on Depth Perception

We pretrain the SimCLR-ResNet on CATDiet using the CO3D training set, followed by training its linear probes on the 3D-PC training set for binary depth-order classification.

		SAY [107] Acc	SAY-C [45] mCE	IN [92] Acc	IN-C [45] mCE	TSCC [35] S-Bias	Sil-O [35] Acc	3D-PC [63] dAcc
S-V [18][31]	**Comb**	**55.1**	**79.0**	**64.1**	**71.4**	**20.7**	**46.7**	**73.2**
	SHF	49.7	85.7	62.8	80.4	13.8	33.3	66.3
	STD	49.0	87.7	64.0	77.3	**20.7**	33.3	60.5
S-R [18][41]	**Comb**	**63.1**	**77.3**	**75.5**	**63.9**	**41.7**	46.7	**70.5**
	SHF	56.4	86.1	72.5	73.3	37.5	50.0	63.5
	STD	55.1	85.8	73.3	71.5	37.5	**56.7**	64.4
D-V [15][31]	**Comb**	**55.1**	**75.7**	65.6	**69.8**	**23.3**	**63.0**	**77.2**
	SHF	52.3	81.2	62.9	77.3	10.0	33.3	74.1
	STD	54.5	79.5	**68.0**	**69.8**	20.0	46.7	73.6
D-R [15][41]	**Comb**	**58.2**	**80.7**	**76.3**	**66.5**	**25.9**	**50.0**	**67.6**
	SHF	43.0	101.1	66.8	88.7	22.2	40.0	63.2
	STD	51.7	92.0	74.7	78.4	22.2	**50.0**	63.6

Table 1. **Performance of SSL models pretrained on CombDiet from the SAY dataset and their baselines on object recognition and depth perception tasks.** Each row corresponds to a specific [SSL]-[backbone] configuration (four in total). In the first column, **S**, **V**, **R**, and **D** denote SimCLR [18], ViT [31], ResNet [41], and DINO [15], respectively; **Comb** indicates our proposed CombDiet. Within each row, three models are compared: CombDiet, SHF, and STD. Columns report results on the respective datasets with their corresponding evaluation metrics. Shaded rows highlight CombDiet performance; best results are shown in **bold**.

The model and its baselines are then evaluated on the 3D-PC test set and the IEVC dataset, which simulates the Visual Cliff paradigm [37].

Progressive depth perception emerges during pretraining with infant-like staged diets. As shown in **Fig. 4b**, the dAcc on the 3D-PC dataset for the SSL model pretrained on CATDiet increases sharply around the fifth epoch, marking the emergence of depth sensitivity. In contrast, CAT-SHF plateaus early with a lower dAcc of 61.5%, indicating that disrupting the developmental sequence impairs the acquisition of depth cues. This result provides a computational account suggesting that pictorial depth sensitivity in infants emerges postnatally rather than being innate [5].

Models pretrained with progressive staged diets align with infant behaviors in the visual cliff paradigm. In the visual cliff evaluation (**Fig. 4c**), the SSL model pretrained on CATDiet correctly predicts the depth order across all three synthesized environments in the IEVC dataset. Specifically, the model identifies the shallow side as safe and avoids the deep side, mirroring the avoidance behavior observed in one-year-old infants [37]. In contrast, CAT-SHF fails to make accurate depth-order predictions, underscoring the importance of developmental order in visual diets. Consistent with [5], these results also suggest that depth perception is not innate but rather emerges through early visual experience.

5.3. CombDiet Generalizes to Real-World Tasks

We first pretrain four SSL models on CombDiet using the SAY training set followed by their linear probes trained on the annotated set for 26-way image classification. The models and their baselines are evaluated on the clean SAY test images and the corrupted SAY-C images for object recognition. Additionally, we train another set of linear probes on IN for 15-way image classification and evaluate them on out-of-domain object recognition datasets, including IN, IN-C, TSCC, and Sil-O. Finally, we train and evaluate their models with separate 2-way linear probes for the binary depth-order classification task on 3D-PC. We provide additional results for CO3D-pretrained SSL models in **Sec. S3**. Similar analyses apply to these models as well.

CombDiet models achieve the best performance among all the baselines in object recognition and depth-order estimation tasks. As shown in Column 1 and Column 9 of **Tab. 1**, models pretrained with CombDiet consistently outperform all baselines across the object recognition and depth-order classification benchmarks on the SAY and 3D-PC datasets. For instance, for SimCLR-ViT, the CombDiet model achieves the highest Acc of 55.1% on the clean SAY test images and the highest dAcc of 73.2% on the 3D-PC dataset. Interestingly, CombDiet-SHF performs comparably to STD, though both remain inferior to our CombDiet models. This indicates that the developmental order in the progressive schedule is critical; disrupting it degrades model performance to the level of STD.

CombDiet models are more robust to recognizing out-of-domain images and exhibit stronger inductive biases towards shapes. Columns 2–8 in **Tab. 1** show model performance on out-of-domain image recognition tasks. CombDiet models are more robust to corrupted images in SAY-C compared to all baselines. Remarkably, they also generalize to IN and IN-C—datasets never seen during training, which demonstrates strong out-of-distribution generalization capabilities. Despite never being exposed to TSCC images or any human behavioral data, CombDiet models exhibit human-like shape biases, which become even more pronounced on Sil-O, where object recognition relies solely on silhouettes. These results indicate that our developmental visual diets induce strong shape-based inductive biases, improving model generalization and practical applicability in real-world vision tasks.

6. Discussion

We present CATDiet, a progressive visual diet for SSL that simulates infant vision through staged constraints on color, acuity, and temporal continuity. Across ten comprehensive benchmarks, CATDiet enhances robustness in object recognition and exhibits biologically aligned developmental patterns, including neural plasticity changes,

early emergence of depth perception, and infant-like hesitation in the visual cliff paradigm. Building on these insights, CombDiet integrates these developmental principles with standard SSL training, achieving superior in-domain and out-of-domain performance on object recognition and depth-order classification tasks. Our work provides an initial step toward modeling the progressive structure of early visual experience and offers a framework for scaffolding robust visual representations in machines. Future work could extend this approach by incorporating biologically inspired mechanisms—such as recurrent connections—to better capture how humans internalize experience and form robust visual representations.

References

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International conference on learning representations*, 2018. 6, 7

[2] Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 37–46, 2001. 3

[3] Michael J Arcaro, Peter F Schade, Justin L Vincent, Carlos R Ponce, and Margaret S Livingstone. Seeing faces is necessary for face-domain formation. *Nature neuroscience*, 20(10):1404–1412, 2017. 1

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3

[5] R. N. Aslin. Visual development: Infant. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 16250–16255. Pergamon, 2001. 1, 8

[6] Arthur Aubret, Céline Teulière, and Jochen Triesch. Self-supervised visual learning from interactions with objects. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024. 3

[7] Vladislav Ayzenberg, Sukran Bahar Sener, Kylee Novick, and Stella F Lourenco. Fast and robust visual object recognition in young children. *Science Advances*, 11(27): eads6821, 2025. 3

[8] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. *Advances in neural information processing systems*, 31, 2018. 1

[9] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 3

[10] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 3

[11] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 3

[12] Colin Blakemore and Grahame F Cooper. Development of the brain depends on the visual environment. *Nature*, 228 (5270):477–478, 1970. 1

[13] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021. 3

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2, 3

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 4, 8, 1

[16] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021. 3

[17] Pengguang Chen, Shu Liu, Hengshuang Zhao, Xingquan Wang, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 1

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2, 3, 4, 8

[19] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 3

[20] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2, 3

[21] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[22] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 3

[23] Ming-Chang Chiu, Yingfei Wang, Derrick Eui Gyu Kim, Pin-Yu Chen, and Xuezhe Ma. On human visual contrast sensitivity and machine vision robustness: A comparative study. *arXiv preprint arXiv:2212.08650*, 1(3), 2022. 1

[24] Elizabeth M Clerkin, Elizabeth Hart, James M Rehg, Chen Yu, and Linda B Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, 2017. 1

[25] Michael A Crognale. Development, maturation, and aging of chromatic visual pathways: Vep results. *Journal of Vision*, 2(6):2–2, 2002. 1, 4

[26] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 1

[27] Tessa M Dekker and Roni O Maimon-Mor. How infants look shapes what they learn. *Proceedings of the National Academy of Sciences*, 122(20):e2505492122, 2025. 1

[28] Karen R Dobkins, Christina M Anderson, and John Kelly. Development of psychophysically-derived detection contours in l-and m-cone contrast space. *Vision Research*, 41(14):1791–1807, 2001. 4

[29] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016. 1

[30] Melissa Dominguez and Robert A Jacobs. Developmental constraints aid the acquisition of binocular disparity sensitivities. *Neural Computation*, 15(1):161–182, 2003. 1

[31] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020. 4, 5, 8, 3

[32] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pages 3015–3024. PMLR, 2021. 2, 3

[33] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3299–3309, 2021. 3

[34] John M Franchak, Linda Smith, and Chen Yu. Developmental changes in how head orientation structures infants' visual attention. *Developmental psychobiology*, 66 (7):e22538, 2024. 1

[35] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018. 2, 5, 8

[36] Sybil Geldart, Catherine J Mondloch, Daphne Maurer, Scania De Schonen, and Henry P Brent. The effect of early visual deprivation on the development of face processing. *Developmental Science*, 5(4):490–501, 2002. 1

[37] Eleanor J Gibson and Richard D Walk. The" visual cliff". *Scientific American*, 202(4):64–71, 1960. 2, 8

[38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3

[39] Manu Srinath Halvagal and Friedemann Zenke. The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *Nature neuroscience*, 26(11):1906–1915, 2023. 4

[40] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33:5679–5690, 2020. 3

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 8, 3

[42] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[43] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 3

[44] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 2, 3

[45] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2, 3, 5, 6, 8

[46] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *International Conference on Learning Representations*, 2019. 1, 3

[47] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 3

[48] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993. 3

[49] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22 (2):85–126, 2004. 3

[50] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 3

[51] Scott P Johnson. How infants learn about the visual world. *Cognitive science*, 34(7):1158–1184, 2010. 1

[52] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings*

of the IEEE/CVF international conference on computer vision, pages 4773–4783, 2019. 3

[53] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9619–9628, 2021. 3

[54] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Self-taught metric learning without labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7431–7441, 2022. 3

[55] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022. 4

[56] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10326–10335, 2021. 3

[57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[58] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. 3

[59] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 3

[60] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. Ffcv: Accelerating training by removing data bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12011–12020, 2023. 1

[61] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 3

[62] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 3

[63] Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. *arXiv preprint arXiv:2406.04138*, 2024. 2, 5, 8, 3

[64] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651*, 2018. 3

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[66] Spandan Madan, Tomotake Sasaki, Tzu-Mao Li, Xavier Boix, and Hanspeter Pfister. Small in-distribution changes in 3d perspective and lighting fool both cnns and transformers. *arXiv preprint arXiv:2106.16198*, 3, 2021. 3

[67] Spandan Madan, You Li, Mengmi Zhang, Hanspeter Pfister, and Gabriel Kreiman. Improving generalization by mimicking the human visual diet. *arXiv preprint arXiv:2206.07802*, 2022. 1

[68] John Maule, Alice E Skelton, and Anna Franklin. The development of color perception and cognition. *Annual Review of Psychology*, 74(1):87–111, 2023. 1, 4

[69] Daphne Maurer, Catherine J Mondloch, and Terri L Lewis. Sleeper effects. *Developmental Science*, 10(1):40–47, 2007. 1, 7

[70] D Luisa Mayer and Velma Dobson. Visual acuity development in infants and young children, as assessed by operant preferential looking. *Vision research*, 22(9): 1141–1151, 1982. 1

[71] Ayelet McKyton, Itay Ben-Zion, Ravid Doron, and Ehud Zohary. The limits of shape recognition following late emergence from blindness. *Current Biology*, 25(18): 2373–2378, 2015. 2

[72] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021. 1, 3

[73] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017. 3

[74] Margaret C Moulson, Robert W Shannon, and Charles A Nelson. Neural correlates of visual recognition in 3-month-old infants: The role of experience. *Developmental psychobiology*, 53(4):416–424, 2011. 1

[75] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013. 3

[76] Masako Myowa-Yamakoshi, Yuka Kawakita, Mako Okanda, and Hideko Takeshita. Visual experience influences 12-month-old infants' perception of goal-directed actions of others. *Developmental psychology*, 47(4):1042, 2011. 1

[77] Quan Nguyen and Koushil Sreenath. Robust safety-critical control for dynamic robotics. *IEEE Transactions on Automatic Control*, 67(3):1073–1088, 2021. 1

[78] Anthony M Norcia and Christopher W Tyler. Spatial frequency sweep vep: visual acuity during the first year of life. *Vision research*, 25(10):1399–1408, 1985. 1, 4

[79] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[80] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 3

[81] A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3): 271–283, 2024. 2, 3

[82] Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33: 9960–9971, 2020. 2, 3

[83] Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765, 2023. 3

[84] Nikhil Parthasarathy, Olivier J Hénaff, and Eero P Simoncelli. Layerwise complexity-matched learning yields an improved model of cortical area v2. *ArXiv*, pages arXiv–2312, 2024. 4

[85] Zachary J Petroff, Swapnaa Jayaraman, Linda B Smith, T Rowan Candy, and Kathryn Bonnen. The world through infant eyes: Evidence for the early emergence of the cardinal orientation bias. *Proceedings of the National Academy of Sciences*, 122(16):e2421277122, 2025. 1

[86] Lisa Putzar, Kirsten Hötting, and Brigitte Röder. Early visual deprivation affects the development of face recognition and of audio-visual speech perception. *Restorative neurology and neuroscience*, 28(2):251–257, 2010. 1

[87] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021. 3

[88] Pasko Rakic, Jean-Pierre Bourgeois, Maryellen F Eckenhoff, Nada Zecevic, and Patricia S Goldman-Rakic. Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science*, 232(4747): 232–235, 1986. 7

[89] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in neural information processing systems*, 34:29935–29948, 2021. 1

[90] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 2, 5, 3

[91] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019. 3

[92] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2, 5, 8, 1, 3

[93] Akira Sakai, Taro Sunagawa, Spandan Madan, Kanata Suzuki, Takashi Katoh, Hiromichi Kobashi, Hanspeter Pfister, Pawan Sinha, Xavier Boix, and Tomotake Sasaki. Three approaches to facilitate invariant neurons and generalization to out-of-distribution orientations and illuminations. *Neural Networks*, 155:119–143, 2022. 3

[94] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv preprint arXiv:1912.12510*, 2019. 3

[95] Felix Schneider, Xia Xu, Markus R Ernst, Zhengyang Yu, and Jochen Triesch. Contrastive learning through time. In *SVRHM 2021 Workshop@ NeurIPS*, 2021. 3

[96] Gudrun Schwarzer. How motor and visual experiences shape infants' visual processing of objects and faces. *Child Development Perspectives*, 8(4):213–217, 2014. 1

[97] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021. 3

[98] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 3

[99] Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36:54199–54212, 2023. 2, 3

[100] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3

[101] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2515, 2024. 1

[102] Parantak Singh, You Li, Ankur Sikarwar, Stan Weixian Lei, Difei Gao, Morgan B Talbot, Ying Sun, Mike Zheng Shou, Gabriel Kreiman, and Mengmi Zhang. Learning to learn: How to continuously teach humans and machines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11719, 2023. 1

[103] Alice E Skelton, John Maule, and Anna Franklin. Infant color perception: Insight into perceptual development. *Child development perspectives*, 16(2):90–95, 2022. 1, 4

[104] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 1

[105] Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 8:296143, 2017.

[106] Linda B Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4): 325–336, 2018. 1

[107] Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29, 2021. 3, 5, 8, 1

[108] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 3

[109] Morgan B Talbot, Rushikesh Zawar, Rohil Badkundri, Mengmi Zhang, and Gabriel Kreiman. Tuned compositional feature replays for efficient stream learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[110] Davida Y Teller. First glances: the vision of infants. the friedenwald lecture. *Investigative ophthalmology & visual science*, 38(11):2183–2203, 1997. 1, 4

[111] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[112] Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13806–13815, 2020. 3

[113] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016. 1

[114] Ramakrishna Vedantam, David Lopez-Paz, and David J Schwab. An empirical investigation of domain generalization with empirical risk minimizers. *Advances in neural information processing systems*, 34:28131–28143, 2021. 3

[115] Lukas Vogelsang, Sharon Gilad-Gutnick, Evan Ehrenberg, Albert Yonas, Sidney Diamond, Richard Held, and Pawan Sinha. Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44): 11333–11338, 2018. 1, 2, 3, 7

[116] Lukas Vogelsang, Marin Vogelsang, Gordon Pipa, Sidney Diamond, and Pawan Sinha. Butterfly effects in perceptual development: A review of the 'adaptive initial degradation'hypothesis. *Developmental Review*, 71: 101117, 2024. 1

[117] Marin Vogelsang, Lukas Vogelsang, Priti Gupta, Tapan K Gandhi, Pragya Shah, Piyush Swami, Sharon Gilad-Gutnick, Shlomit Ben-Ami, Sidney Diamond, Suma Ganesh, et al. Impact of early visual experience on later usage of color cues. *Science*, 384(6698):907–912, 2024. 1, 2, 4, 7

[118] Marin Vogelsang, Lukas Vogelsang, Gordon Pipa, Sidney Diamond, and Pawan Sinha. Potential role of developmental experience in the emergence of the parvo-magno distinction. *Communications Biology*, 8(1): 987, 2025. 1, 2, 3

[119] Alex N Wang, Christopher Hoang, Yuwen Xiong, Yann LeCun, and Mengye Ren. Poodle: Pooled and dense self-supervised learning from naturalistic videos. *arXiv preprint arXiv:2408.11208*, 2024. 3

[120] Chuyao Wang and Nabil Aouf. Explainable deep adversarial reinforcement learning approach for robust autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024. 1

[121] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6678–6687, 2020. 3

[122] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 3

[123] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A survey on the robustness of computer vision models against common corruptions. *arXiv preprint arXiv:2305.06024*, 2023. 1

[124] Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763. IEEE, 2021. 3

[125] Ziyu Wang, Shuangpeng Han, and Mengmi Zhang. Pose prior learner: Unsupervised categorical prior learning for pose estimation. *arXiv preprint arXiv:2410.03858*, 2024. 1

[126] John S Werner and BR Wooten. Human infant color vision and color perception. *Infant Behavior and Development*, 2: 241–273, 1979. 1, 4

[127] Justin N Wood and Samantha MW Wood. The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4): 1391–1406, 2018. 1, 4

[128] Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10149–10159, 2021. 3

[129] Jay Zhangjie Wu, David Junhao Zhang, Wynne Hsu, Mengmi Zhang, and Mike Zheng Shou. Label-efficient online continual object detection in streaming video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19246–19255, 2023. 1

[130] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance

discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[131] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[132] Xiaohao Xu, Tianyi Zhang, Sibo Wang, Xiang Li, Yongqi Chen, Ye Li, Bhiksha Raj, Matthew Johnson-Roberson, and Xiaonan Huang. From perfect to noisy world simulation: Customizable embodied multi-modal perturbations for slam robustness benchmarking. *arXiv preprint arXiv:2406.16850*, 2024. 1

[133] Takuto Yamamoto, Hirosato Akahoshi, and Shigeru Kitazawa. Emergence of human-like attention and distinct head clusters in self-supervised vision transformers: A comparative eye-tracking study. *Neural Networks*, page 107595, 2025. 4

[134] Xiao Yang, Lingxuan Wu, Lizhong Wang, Chengyang Ying, Hang Su, and Jun Zhu. Reinforced embodied active defense: Exploiting adaptive interaction for robust visual perception in adversarial 3d environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[135] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6210–6219, 2019. 3

[136] Thomas Yerxa, Jenelle Feather, Eero Simoncelli, and SueYeon Chung. Contrastive-equivariant self-supervised learning improves alignment with primate visual area it. *Advances in neural information processing systems*, 37: 96045–96070, 2024. 4

[137] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1, 3

[138] Lorijn Zaadnoordijk, Tarek R Besold, and Rhodri Cusack. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520, 2022. 1

[139] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 3

[140] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019. 3

[141] Arthur Zhang, Chaitanya Eranki, Christina Zhang, Ji-Hwan Park, Raymond Hong, Pranav Kalyani, Lochana Kalyanaraman, Arsh Gamare, Arnav Bagad, Maria Esteva, et al. Toward robust robot 3-d perception in urban environments: The ut campus object dataset. *IEEE Transactions on Robotics*, 40:3322–3340, 2024. 1

[142] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2017. 1, 3

[143] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adversarial relighting against face recognition. *IEEE Transactions on Information Forensics and Security*, 19:9145–9157, 2024. 3

[144] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 3

[145] Ruixin Zhao, Sai Hong Tang, Jiazheng Shen, Eris Elianddy Bin Supeni, and Sharafiz Abdul Rahim. Enhancing autonomous driving safety: A robust traffic sign detection and recognition model tsd-yolo. *Signal Processing*, 225:109619, 2024. 1

[146] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1

[147] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. 4

Supplementary Material for Learning to See Through a Baby's Eyes: Early Visual Diets Enable Robust Visual Intelligence in Humans and Machines

S1. Implementation Details

Training Details.

CATDiet integrates the staged curricula of CDiet and ADiet by interleaving their respective schedules, yielding an eight-stage training curriculum with stage lengths of [10, 6, 1, 5, 1, 2, 3, 2] epochs. Across these stages, we jointly manipulate the Gaussian blur standard deviation σ and color-saturation ratio s. Specifically, we set $\sigma = [4, 3, 2, 2, 1, 1, 0, 0]$ (in pixels) and assign the saturation ranges for stages one through eight as (0.20, 0.36), (0.36, 0.52), (0.36, 0.52), (0.52, 0.68), (0.52, 0.68), (0.68, 0.84), (0.68, 0.84), and (0.84, 1.0), respectively.

We accelerate data loading using FFCV [60]. Training hyperparameters are selected via grid search for optimal performance. For SimCLR, we employ a staged temperature schedule; the temperature values τ for CDiet, ADiet, TDiet, CATDiet, and CombDiet are listed in **Tab. S1**.

	τ value across stages	stage durations
Cdiet	[0.5, 0.4, 0.3, 0.2, 0.1]	[10, 7, 6, 5, 2]
Adiet	[0.5, 0.4, 0.3, 0.2, 0.1]	[10, 6, 6, 3, 5]
Tdiet	[0.1]	[30]
CATDiet	[0.5, 0.45, 0.4, 0.35, 0.3, 0.2, 0.15, 0.1]	[10, 6, 1, 5, 1, 2, 3, 2]
CombDiet	[0.5, 0.45, 0.4, 0.35, 0.3, 0.2, 0.15, 0.1, 0.1]	[10, 6, 1, 5, 1, 2, 3, 2, 70]

Table S1. **Temperature (τ) schedules used for SimCLR under different visual diets.** Each row corresponds to one visual diet. The second column specifies the τ value used in each stage, and the third column provides the duration of each stage (in epochs).

For DINO, temperature is fixed throughout training; we use a student temperature of $\tau_s = 0.1$ and a teacher temperature of $\tau_t = 0.04$ for all visual diets. For momentum, we use the update schedule from [15] for CATDiet and each individual diet (CDiet, ADiet and TDiet); in the two-phase CombDiet setting, the same update schedule is applied in Phase 1 and the momentum is reinitialized at the beginning of Phase 2 to adapt to SDiet.

Calculation of mCE. We compute the mean Corruption Error (mCE) following the protocol of [45]. ImageNet-C comprises 15 corruption types, each evaluated at 5 severity levels. For a classifier f, let $E^f_{s,c}$ denote the top-1 error under corruption type c at severity level s. The aggregated error for corruption type c is $E^f_c = \sum_{s=1}^{5} E^f_{s,c}$.

Because corruption types differ in difficulty, we normalize these errors by those of a baseline model, E^{baseline}_c. Following [45], we use AlexNet [57] as the reference, yielding the normalized corruption error:

$$CE^f_c = \frac{E^f_c}{E^{\text{AlexNet}}_c}$$

The mCE is then obtained by averaging CE^f_c over all 15 corruption types, providing an overall measure of corruption robustness for model f.

Mapping Classes between IN and SAY. The 15 IN [92] classes used in this study and their corresponding SAY [107] labels are listed below, where each IN label is followed by its corresponding SAY label in parentheses: n02124075 (cat), n09421951 (sand), n04399382 (plushanimal), n04590129 (window), n04239074 (door), n03125729 (crib), n02802426 (ball), n03201208 (table), n04344873 (couch), n03180011 (computer), n04099969 (chair), n03598930 (puzzle), n04285008 (car), n04204238 (basket), and n04462240 (toy).

S2. Object recognition performance of CATDiet on other models

As shown in **Fig. S1**, the conclusions in **Sec. 5.1** consistently generalize across all SSL methods and backbones: Across all models (i.e., across all panels in **Fig. S1**), each individual diet (CDiet, ADiet, TDiet) consistently achieves lower mCE than its corresponding baselines, demonstrating that developmentally-inspired visual diets promote robust representation learning across diverse SSL configurations. Moreover, the integrated CATDiet not only attains the lowest mCE among all

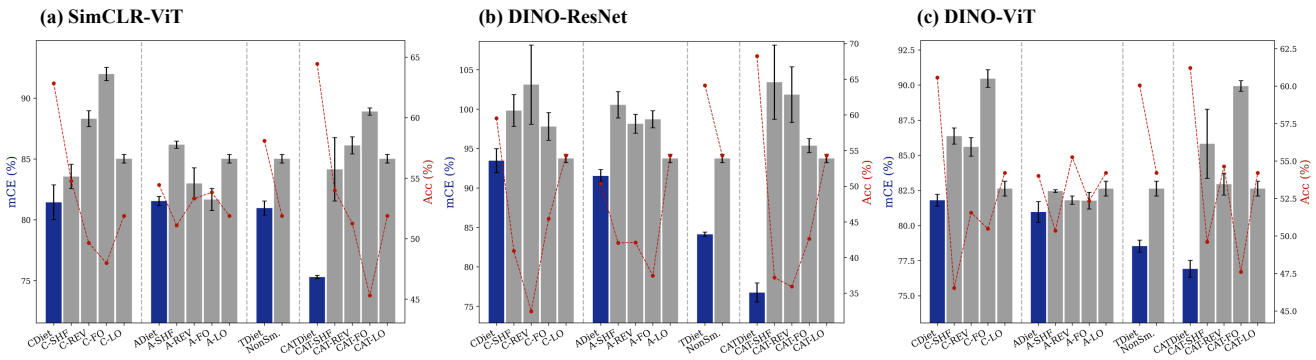

Figure S1. **Object recognition performance on CO3D (clean) and CO3D-C (corrupted) datasets for three SSL models pretrained on CATDiet and its individual diets.** Panels (a–c) show results for SimCLR-ViT, DINO-ResNet, and DINO-ViT, respectively. Within each panel, bars show mCE (↓, left axis) for our diets (blue) and their baselines (gray); the dashed red line indicates Acc (↑, right axis). Error bars reflect the standard error of mCE over three runs. Within each panel, the four groups correspond to different attributes of the proposed visual diets: Color, Acuity, Temporality, and their combination.

diets but also exhibits larger performance gains in mCE and Acc over its baselines than any individual diet across all models, confirming that the synergistic benefit of integrating all developmentally-inspired diets generalize across other SSL methods and backbones.

S3. Object Recognition and Depth Order Classification Results of CombDiet pretrained on CO3D-10 and CO3D-50

Sec. 5.3 demonstrates that CombDiet improves model robustness and achieves superior performance on both object recognition and depth-order classification tasks. To assess whether these gains extend beyond SAY, we replicate the same experiments on CO3D. We begin with the 10-class CO3D subset used in the main text, and then evaluate scalability by repeating the experiments on an expanded 50-class version, which we denote CO3D-50.

Datasets.

CO3D-50 [90] and **CO3D-50-C**. We construct the **CO3D-50** benchmark using all 50 object categories in the CO3D dataset, sampling approximately 26,000 object instances for our experiments. Training and test splits are defined at the object-instance level within each category, and details of the frame-sampling procedure for both splits are provided in **Sec. 4.1**. To evaluate robustness, we additionally construct a corrupted version of the CO3D-50 test set following the ImageNet-C protocol [45], resulting in **CO3D-50-C**. For clarity, the 10-category subset used in the main text and its corrupted counterpart are referred to throughout as **CO3D-10** and **CO3D-10-C**, respectively.

IN-10, **IN-40**, **IN-10-C**, and **IN-40-C** [92]. For out-of-domain object recognition, we construct two ImageNet-21K subsets based on category overlap with CO3D-10 and CO3D-50. For CO3D-10, the 10 overlapping ImageNet-21K categories form **IN-10**. For CO3D-50, 40 categories align with ImageNet-21K, forming **IN-40**. To evaluate robustness, we generate corrupted versions of the IN-10 and IN-40 test sets using the ImageNet-C protocol [45], producing **IN-10-C** and **IN-40-C**. Complete class mappings between CO3D-10 and IN-10, and between CO3D-50 and IN-40, are listed below, where each ImageNet label is followed by its corresponding CO3D class in parentheses:

The overlapping classes between IN-10 and CO3D-10 are: n07930864 (cup), n04146614 (toybus), n03791053 (motorcycle), n04399382 (teddybear), n03809312 (toyplane), n04026813 (bicycle), n04099969 (chair), n03417042 (toytruck), n02958343 (car), and n02971579 (toytrain).

The overlapping classes between IN-40 and CO3D-50 are: n07930864 (cup), n04146614 (toybus), n03791053 (motorcycle), n04399382 (teddybear), n04552348 (toyplane), n02835271 (bicycle), n04099969 (chair), n03417042 (toytruck), n04285008 (car), n04335435 (toytrain), n03483316 (hairdryer), n04522168 (vase), n04263257 (bowl), n04507155 (umbrella), n03891332 (parkingmeter), n04442312 (toaster), n04447861 (toilet), n02992529 (cellphone), n07248320 (book), n02802426 (ball), n04344873 (couch), n03983396 (bottle), n03991062 (plant), n03085013 (keyboard), n03793489 (mouse), n01608432 (kite), n07873807 (pizza), n02823750 (wineglass), n02769748 (backpack), n07747607 (orange), n03709823 (handbag), n07753592 (banana), n04404412 (tv), n04074963 (remote), n07742313 (apple), n07697537 (hotdog), n07714990 (broccoli), n03891251 (bench), n03642806 (laptop), and n03761084 (microwave).

Detailed Settings. We evaluate CombDiet models pretrained on CO3D-10 and CO3D-50 across object recognition and depth-order classification. We summarize the specific training, linear probing, and test procedures below.

(1) *Pretrained on CO3D-10.* We pretrain four SSL models with CombDiet on the CO3D-10 training set. For object recognition, we train 10-way linear probes on CO3D-10 and evaluate them on both clean CO3D-10 test images and the corrupted CO3D-10-C set. To assess out-of-domain generalization, we train 10-way probes on IN-10 and evaluate on IN-10 and IN-10-C. Depth perception is measured using a 2-way linear probe trained and tested on the 3D-PC Depth Order dataset.

(2) *Pretrained on CO3D-50.* We pretrain four SSL models with CombDiet on the CO3D-50 training set. For object recognition, we train 50-way probes and evaluate them on CO3D-50 and CO3D-50-C. For out-of-domain evaluation, we train 40-way probes on IN-40 and evaluate on IN-40 and IN-40-C. Depth perception is again evaluated using a 2-way linear probe trained and tested on the 3D-PC dataset.

Results. Following the presentation style in **Sec. 5.3**, we report results in **Tab. S2** for both CO3D-10 (Columns 3–7) and CO3D-50 (Columns 8–12). On CO3D-10, CombDiet models achieve competitive or superior performance in Acc and dAcc, and outperform all baselines by a substantial margin in mCE. We observe similar trends on CO3D-50, which includes a much broader set of object categories. These results demonstrate that the progressive ordering of developmental diets is essential: it substantially improves model robustness, whereas disrupting this order reduces performance to that of standard SSL training.

		CO3D					CO3D-50				
		CO3D-10 [90] Acc	CO3D-10-C [45] mCE	IN-10 [92] Acc	IN-10-C [45] mCE	3D-PC [63] dAcc	CO3D-50 [90] Acc	CO3D-50-C [45] mCE	IN-40 [92] Acc	IN-40-C [45] mCE	3D-PC [63] dAcc
SimCLR-ViT [18] [31]	***CombDiet***	74.4	**67.6**	**80.4**	**57.8**	**74.5**	**75.2**	**76.6**	**65.1**	**75.4**	**71.3**
	SHF	74.0	69.8	77.6	61.2	71.1	71.1	83.7	61.7	81.4	65.7
	STD	**74.8**	68.6	78.4	58.7	69.3	71.8	83.5	61.5	82.2	61.3
SimCLR-ResNet [18] [41]	***CombDiet***	**83.2**	**58.7**	**88.8**	**44.8**	**69.0**	**82.4**	**70.5**	**76.0**	**69.1**	**72.4**
	SHF	82.3	64.6	86.2	51.1	64.0	79.7	80.1	74.8	76.2	65.3
	STD	80.7	63.2	88.0	51.8	62.8	78.7	84.1	74.0	77.4	68.8
DINO-ViT [15] [31]	***CombDiet***	**77.7**	**62.8**	81.8	**54.8**	**79.9**	80.0	**67.9**	71.8	**66.0**	**78.2**
	SHF	73.8	68.7	82.0	63.2	76.3	12.1	120.7	9.4	122.2	52.5
	STD	76.3	64.0	**82.6**	57.2	78.5	**80.8**	70.9	**73.7**	67.7	77.4
DINO-ResNet [15] [41]	***CombDiet***	83.5	**57.1**	**89.6**	**48.9**	**70.9**	80.0	**74.4**	77.2	**72.8**	72.8
	SHF	78.0	74.7	85.4	69.1	63.4	57.6	102.6	57.2	100.5	58.2
	STD	**85.3**	64.5	89.2	59.8	68.4	**85.2**	75.4	**79.7**	74.5	**75.0**

Table S2. **Object Recognition and depth-order classification results of SSL models pretrained with CombDiet on CO3D-10 and CO3D-50 respectively.** Columns 3-7 denote results of models pretrained on CO3D-10, columns 8-12 denote results of models pretrained on CO3D-50. Each row corresponds to a specific [SSL]-[backbone] configuration (four in total). Within each row, three models are compared: CombDiet, SHF, and STD. Shaded rows highlight CombDiet performance; best results are shown in **bold**.