A RELATED WORK

Our work connects most closely to human-in-the-loop data augmentation and the expansive literature surrounding human categorical perception from the cognitive science community, as well as ongoing efforts in the machine learning community to develop more efficacious mixup-based data and label mixing functions.

A.1 HUMAN-IN-THE-LOOP DATA AUGMENTATION

Incorporating expert feedback into the learning procedure has received increasing attention [Chen et al., 2022]. In particular, previous work has considered incorporating humans “in the loop” for data augmentation. For instance, DatasetGAN [Zhang et al., 2021] employs human participants to label GAN-generated images and feeds these back to the model to generate more synthetic data. [Kaushik et al., 2019] similarly incorporate human feedback by having humans create counterfactual samples, and has been shown to be an efficient method to adjust model behavior [Kaushik et al., 2021]. Other works have considered employing humans to provide “rationales” about examples to improve data-efficiency and downstream modeling performance [Zaidan et al., 2007]. Here, we marry these ideas in the context of mixup by eliciting data and label-mixing function parameters to align with human percepts.

A.2 HUMAN CATEGORICAL PERCEPTION

In cognitive science, eliciting humans’ judgments over synthetically-constructed examples is a tried-and-true method to characterize human category boundaries [Newell and Bülthoff, 2002, Folstein et al., 2013, Feldman, 2021, Folstein et al., 2012]. Such studies often reveal a non-linear structure of humans’ percepts. For instance, in the audio domain, the identification of vowel categories has been found to demonstrate “warping” close to prototypical category members – known as the “perceptual magnet effect” [Kuhl, 1991, Feldman et al., 2009]. Similar nonlinearities have been found in the perception of boundaries between face identities [Beale and Keil, 1995] and the transitions between 3D shapes [Newell and Bülthoff, 2002, Destler et al., 2019]. Our linearly interpolated stimuli are similar in spirit to the morphological trajectories used in these works, as well as other synthetically-combined images [Oliva et al., 2006]. [Gruber et al., 2018] also consider 50/50 mixed images; however, their elicitation involves open-ended judgments which does not permit the same kind of data and label mixing alignment studies as our methods more directly elicit human-inferred generative parameters. Our work also connects to other non-linear perceptual phenomena encountered in the visual domain; namely, binocular rivalry, whereby present participants with a different image in each eye has been shown to induce oscillatory percepts [Blake and Logothetis, 2002, Tong et al., 2006].

*Correspondence to: kmc61@cam.ac.uk

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).
A.3 OTHER MIXUP-BASED SYNTHETIC DATA SCHEMES

Many alternative mixup data and label mixing functions have been proposed [Verma et al., 2019, Yun et al., 2019, Kim et al., 2020a,b, Hendrycks et al., 2022]. Closest to our work, [Sohn et al., 2022] highlight particular issues with the linear interpolation in label space on the learned topology of the model’s category boundaries and instead utilize a Gaussian Mixture Model (GMM)-based relabeling scheme to construct “better” labels than those used in baseline mixup. Additional work on learning better pseudo-labels over mixup samples have been proposed [Arazo et al., 2020, Cascante-Bonilla et al., 2020, Sohn et al., 2020, Qiu et al., 2022]. Similarly, Between-class (BC) learning [Tokozume et al., 2017, 2018] proposes hand-crafted adjustments to label construction to better align with human perception based on waveform modulations; however, to our knowledge, no previous works have directly considered incorporating humans in-the-loop for either the construction of mixup samples or associated relabeling.

B ADDITIONAL NOTES ON H-MIX

B.1 HUMAN SUBJECT EXPERIMENTS

We include additional details on our human elicitation studies. For all experiments, we require participants speak English as a first-language and reside in the United States. Across all experiments, the mean age for participants was 37.5 years old (± standard deviation of 12.7 yrs). The self-reported sex breakdown was approximately 57% male and 43% female.

Elicitation (RQ1) Each participant sees a total of 32 mixed images, where the final two are repeats. Repeats are primarily used here to measure raters’ internal consistency. The median time taken per participant per image as 9.30 and 11.01 seconds for the Construct and Select-Shuffled interfaces, respectively. A bonus was offered to encourage participants to provide responses which would match what other participants would provide; we applied this bonus to all participants post-hoc resulting in the average participant being paid at a rate of $11.78.

Multiple Interface Styles (RQ1) Why do we consider two styles of elicitation interfaces? We reason that the first interface could be prone to ordering effects – an astute participant could realize that they can count out where the midpoint is located. This led us to design the second variety (Select-Shuffled) wherein the participant sees all images shuffled simultaneously. We hypothesize that Construct could induce responses biased by the participant’s starting position. To probe this, we run two sub-variants wherein participants start from either \( \lambda_f = 0 \) or \( \lambda_f = 0.9 \).

Elicitation (RQ2) Each participant sees 59 – 62 images, where two images are repeated. Repeats are placed at the end and correspond to the images presented on trials 15 and 20, respectively. The order of the images presented in a batch, as well as the order of the endpoint labels displayed for a given image, are shuffled across participants. We follow the same third-person perspective prompting in Section 3 from [Chung et al., 2019]. Participants are asked “what combinations of classes” they thought other participants would say is “used to make” each image, and “how confident” they thought other participants would be in their estimate. Responses are indicated on a slider per question. An example survey screen can be seen in Fig. 4. Subjects took a median of 8.41 seconds per image and were payed at a rate of $8/hr, with an optional bonus which sought to encourage participants to provide calibrated confidence estimates, similar to that of [Vodrahalli et al., 2021]; the bonus was applied to all participants post-hoc. Each mixed image was seen by at least two different participants each. Our interface is depicted in Fig. 4.

B.2 BREAK FROM MONOTONICITY

For users of H-Mix, it is worth noting that we do encounter some breaks with monotonicity (see Fig. 1) in a few of the aggregated “category boundaries.” We reason this could be in part due to several aspects of our set-up. First, our study involved irregular sampling across the space of mixing coefficients we consider: the 50/50 point is enriched. We ran two phases of elicitation: in the first, we sampled 6 image classes per pair to be shown for three mixing coefficients: 0.5, and one chosen randomly from each of the sets \( \{0.1, 0.25\} \) and \( \{0.75, 0.9\} \), respectively (810 images of the 2070). All 1260

---

1Participants’ selections, for each interface type, change by a median of 0.1 in repeat trials, suggesting some inconsistencies in participants’ judgments which persists across elicitation method.

2We observe a median difference of 0.03 and 0.05 in the inferred mixing coefficient and confidence on repeat trials, indicative of high intra-annotator consistency.
other images are shown for a single mixing coefficient sampled uniformly from the set. Second, while we have human judgments for over 2000 total images, there are less than 50 synthetic images considered for each category pair, giving any participant noise – or the odd image – greater leverage to impact trends. We encourage others to use HILL-MixE Suite and continue to scale this work and elucidate the stability of the inferred mixing coefficient category boundaries we begin to hint at here.

![Figure 1: Category boundary elicited from human participants involves a break with monotonicity.](image)

**C  CONFIDENCE-BASED SMOOTHING DETAILS**

We include further details of our methodology for leveraging human-provided confidence to construct $\hat{y}$ introduced in Section 5. Human-derived soft labels have been demonstrated to be valuable for learning [Nguyen et al., 2013, Peterson et al., 2019, Collins et al., 2022, Sanders et al., 2022]. We transform humans’ reported confidence into a smoothing parameter to induce softness using an exponentially-decaying function of human-provided confidence $\omega$: $a + (b^\omega)$; here, $a = 50$, $b = 0.0001$. We use the transformed confidence for additive smoothing on the two-category $\hat{y}$, spread mass accordingly across the full gamut of classes. That is, we use smooth the mass between a completely uniform distribution and a “two-hot” label which uses the human-derived relabeling. Parameters $a, b$ are selected using a held-out set of regular CIFAR-10 images (from $a \in \{5, 10, 15, 25, 50, 100\}$, $b \in \{0.00001, 0.0001, 0.001, 0.01, 0.1\}$). We recommend the consideration of alternate smoothing functions, which could, for instance, account for miscalibration in humans’ reported confidence.

Further, we compare the impact of learning with aggregated versus de-aggregated participants’ predictions. In Section 5, we considered learning with relabelings averaged across participants for a mixed image, and smoothed with confidence reports averaged across participants. Here, we consider instead separating out participants’ responses to learn with individual relabelings smoothed by individual confidence, closely related to [Wei et al., 2022]. We find in Table 1 that learning with de-aggregated data could potentially offer greater performance gains. However, as [Wei et al., 2022] discuss: whether to aggregate can depend on many factors. Our empirical findings support the need for tailoring label construction in context.
Table 1: Varying whether to aggregate when using incorporating human confidence $\omega$ in label construction.

<table>
<thead>
<tr>
<th>Label Type</th>
<th>CE $\pm$</th>
<th>FGSM $\pm$</th>
<th>Calib $\pm$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (Avg with $\omega$)</td>
<td>1.48 $\pm$ 0.06</td>
<td>8.89 $\pm$ 1.59</td>
<td>0.19 $\pm$ 0.01</td>
</tr>
<tr>
<td>Ours (Separated with $\omega$)</td>
<td>1.44 $\pm$ 0.11</td>
<td>8.33 $\pm$ 1.92</td>
<td>0.19 $\pm$ 0.01</td>
</tr>
</tbody>
</table>

D INTERFACES INCLUDED IN HILL MIXE SUITE

We display sample pages of the interfaces created and used in this work, which we release as part of HILL MixE Suite. Interfaces for Section 3 are shown in Figs. 2 and 3; the interface used Sections 4 is depicted in Fig. 4.

Figure 2: Construct interface where participants press arrow keys to select $\tilde{x}$.

Figure 3: Interface for the selection of a given $\lambda_g$ from a set of possible mixed images.
E ALTERNATIVE SYNTHETIC EXAMPLE CATEGORY COMPOSITION ELICITATION

Given human participants are uncertain about the underlying mixing coefficient in a number of cases, we consider whether the category composition typically used in mixup – e.g., placing mass only on the labels of the images used to form the synthetic combined sample – are reasonable. As demonstrated in the main text and in Fig. 5, a synthetic mixup image may look like something else entirely.

We therefore consider a follow-up small-scale human elicitation study wherein we relax the mixup assumption that the label mixing function must output a label constructed only from the two classes used to form the mixed image – and instead collect $\tilde{y}$ directly by showing the mixed image to human annotators in the form of soft labels. This provides a comparison to the previous human-annotated endpoint label mixing coefficients, and can further inspire useful designs for the label mixing policy.

E.1 STUDY DESIGN

We recruit $N = 8$ participants again from Prolific [Palan and Schitter, 2018], yielding soft labels over a total of 100 mixed images. Each participant saw 25 mixed images; each mixed image of the 100 was seen by two participants. The images are drawn from the same set of stimuli created in Section 4; however, here, we only show images with a mixing coefficient $\in \{0.25, 0.5, 0.75\}$. Participants are told that images are formed by combining other images, and are asked to provide what they think others would see in the image. Participants are asked to specify what others would view as the most probable
E.2 ANALYZING ELICITED SOFT LABELS FOR SYNTHETIC IMAGES

We explore the correspondence between the elicited category compositions of the mixed images with the labels that would be used to generate the mixed image (as would be used in traditional mixup; i.e., placing mass only on two categories). While participants did tend to place probability mass on the generating endpoints that correlated with the mixing coefficient used (Pearson $r = 0.52$), interestingly, we find that participants report thinking that 38.3% ($\pm 0.6\%$) of the probability mass of a label should be placed on different classes from those which are used to create the image. This is remarkable and suggests that mixed images do not consistently look like the labels used to create them, corroborate similar trends found in [Gruber et al., 2018] wherein humans endorse categories which are not present in the image. Hence, alternative labelings even beyond the kind we explore in the main text may be preferred which are more aligned with human percepts. Examples of such labeled mixed images are shown in Fig. 5 and the main text.

Takeaways The typical two-category labels used in mixup do not consistently match human perception. We find that human annotators often assign probabilities to alternate classes when asked to label a mixed image. This suggests that the pursuit of aligning synthetic data labeling to match human perception, at least for the synthetic data constructor used in mixup, warrants the design of alternative label mixing functions $g_{\text{rich}}$ which yield richer label distributions over a broader range of categories.
References


John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. Efficient elicitation approaches to estimate collective crowd answers. In CSCW, 2019. 2, 6

Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In HCOMP, 2022. 3, 6


Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary C. Lipton. Explaining the efficacy of counterfactually-augmented data. ICLR, 2021. 1


Patricia K Kuhl. Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Perception & psychophysics, 50(2):93–107, 1991. 1


Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. Ambiguous images with human judgments for robust visual event classification. In *NeurIPS*, 2022. 3


S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2
