Learning neural decoders without labels using multiple data streams

Journal of Neural Engineering

PAPER

Learning neural decoders without labels using multiple data streams

Steven M Peterson[1,2], Rajesh P N Rao[3,4,5] and Bingni W Brunton[1,2,*]

[1] Department of Biology, University of Washington, Seattle, WA 98195, United States of America
[2] eScience Institute, University of Washington, Seattle, WA 98195, United States of America
[3] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, United States of America
[4] Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195, United States of America
[5] Center for Neurotechnology, University of Washington, Seattle, WA 98195, United States of America
* Author to whom any correspondence should be addressed.

E-mail: bbrunton@uw.edu

Keywords: neural decoding, self-supervised learning, cross-modal learning, deep clustering, electroencephalography

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
1 May 2022

REVISED
13 July 2022

ACCEPTED FOR PUBLICATION
29 July 2022

PUBLISHED
10 August 2022

Abstract

Objective. Recent advances in neural decoding have accelerated the development of brain–computer interfaces aimed at assisting users with everyday tasks such as speaking, walking, and manipulating objects. However, current approaches for training neural decoders commonly require large quantities of labeled data, which can be laborious or infeasible to obtain in real-world settings. Alternatively, self-supervised models that share self-generated pseudo-labels between two data streams have shown exceptional performance on unlabeled audio and video data, but it remains unclear how well they extend to neural decoding. *Approach.* We learn neural decoders without labels by leveraging multiple simultaneously recorded data streams, including neural, kinematic, and physiological signals. Specifically, we apply cross-modal, self-supervised deep clustering to train decoders that can classify movements from brain recordings. After training, we then isolate the decoders for each input data stream and compare the accuracy of decoders trained using cross-modal deep clustering against supervised and unimodal, self-supervised models. *Main results.* We find that sharing pseudo-labels between two data streams during training substantially increases decoding performance compared to unimodal, self-supervised models, with accuracies approaching those of supervised decoders trained on labeled data. Next, we extend cross-modal decoder training to three or more modalities, achieving state-of-the-art neural decoding accuracy that matches or slightly exceeds the performance of supervised models. *Significance.* We demonstrate that cross-modal, self-supervised decoding can be applied to train neural decoders when few or no labels are available and extend the cross-modal framework to share information among three or more data streams, further improving self-supervised training.

1. Introduction

Brain–computer interfaces that decode neural activity to control robotic or virtual devices have shown great potential to assist patients with neurological disabilities [1–7], while also furthering our understanding of brain function [8–10]. Much of the recent progress in brain–computer interfaces has been driven by advances in neural decoding algorithms [11–14]. However, these algorithms typically rely on supervised learning and thus require large amounts of labeled training data; even as large quantities of neural data are now routinely recorded, generating annotated datasets by curating 'ground truth' labels can be laborious and may involve human error [15, 16]. Furthermore, decoding algorithms that are useful in the real world must be able to adapt to new scenarios and non-stationary neural signals with few or no labels [17–20]. One promising approach is to use *self-supervised learning* techniques, which generate pseudo-labels from the data itself and then use those pseudo-labels to train a model iteratively

without prior labels (see [21, 22] for comprehensive reviews of self-supervised approaches). Self-supervised neural decoding models would ideally avoid overfitting to irrelevant variations in the training data and achieve performance comparable to supervised decoders. Such robust, self-supervised neural decoders would largely eliminate the need for tedious data annotation [23, 24], expedite analyses of large, complex neural datasets, and lay the groundwork for brain–computer interfaces that can dynamically recalibrate in real-world settings.

Self-supervised learning has been most successfully applied in natural language processing and computer vision, with performance similar to that of top supervised models. State-of-the-art techniques in natural language processing are often self-supervised [27–29], using held-out words from the training set to learn robust language models. In computer vision, popular self-supervised approaches include learning low-dimensional representations that reconstruct the original input (e.g. autoencoders [30]) as well as contrastive approaches that learn when augmented image pairs are similar or different (e.g. generaive adversarial networks (GANs) and Siamese neural networks [31–34]). However, autoencoders typically minimize the mean squared error between the input and reconstruction, which ignores low amplitude, high frequency activity that may be important for the decoding task [35]. For contrastive approaches, creating dissimilar training pairs can be difficult because there are many ways that an image or signal of interest can differ (e.g. comparing an image of a cat's face with an image of the cat's paw, the back of the cat's head, or a dog's face); in other words, it may not be clear which differences are useful for training a particular model. Several non-contrastive methods have recently been developed that do not require generating dissimilar training pairs [36–39]. For example, deep clustering generates pseudo-labels based on the structure of the data itself, which are then used to iteratively train the model and update the pseudo-labels for the next training step [38]. Many self-supervised techniques have approached or exceeded supervised model performance [33, 36, 37], demonstrating that labeled data is not always necessary to train a robust model.

Even so, self-supervised neural decoding remains a formidable challenge for multiple reasons. First, neural oscillations recorded with scalp electroencephalography (EEG) or intracranial electrocorticography (ECoG) differ greatly from language and image data [40–44]. Relevant features for neural decoding often lack a clear baseline [20, 45], are non-stationary over long time periods [46, 47], exponentially decrease in amplitude at higher frequencies [48, 49], and occur in a small fraction of the total recording electrodes [50, 51]. Second, while contrastive techniques have been applied for neural decoding [35, 52–55], neural data can be noisy and variable from one example to the next, so creating dissimilar examples is often difficult, even with labeled data [56, 57]. Furthermore, many self-supervised approaches in computer vision augment the input images during training to improve model robustness [39, 58, 59]. While random crops, rotations, and translations make sense for image data, deciding how to augment neural data is less clear, especially when the behaviorally relevant frequencies are not known [52, 60] (but see [61]). For these reasons, it can be preferable to use a self-supervised approach for neural decoding that does not rely on contrastive learning and data augmentations.

Sharing information across different data streams provides an intriguing opportunity for self-supervised training of neural decoders. Neuroscience research studies often include multiple data streams recorded simultaneously with the neural recordings, such as muscle activation [62, 63], kinematics of human movements [64–67], and various physiological signals [68, 69]. Each data stream has its own variability, but behaviorally relevant activity should be evident in multiple data streams, assuming each one is acquired at a similar timescale and contains information related to the behaviors of interest. Therefore, sharing information across multiple data streams plausibly provides realistic variations for robust training without requiring data augmentation. Indeed, multi-modal and sensor fusion approaches are already commonly applied to improve supervised neural decoding performance [24, 70–72]. Note that our ultimate goal is to learn neural decoders that make predictions from neural data alone once trained; thus, we cannot simply train decoders using concatenated multi-modal data. For self-supervised learning, Alwassel *et al* [26] showed that combining audio and video data streams improved the performance of non-contrastive, deep clustering models, with performance matching those of supervised models. However, it is unclear how well this cross-modal approach extends to self-supervised neural decoding.

In this paper, we show that cross-modal, self-supervised training yields state-of-the-art neural decoders that approach supervised decoding performance despite only training on unlabeled data (figure 1). We also extend the approach of Alwassel *et al* [26] to any number of data streams and assess the performance of neural decoders trained with this cross-modal, self-supervised approach. We compare cross-modal decoders learned without labels to supervised models and unimodal, self-supervised models. We find that sharing pseudo-labels across multiple data streams substantially improves the performance of self-supervised neural decoding models, with accuracies approaching those of supervised models. When we increase the number of data streams from two to three, cross-modal model performance matches or even slightly exceeds supervised model accuracy. This cross-modal, self-supervised

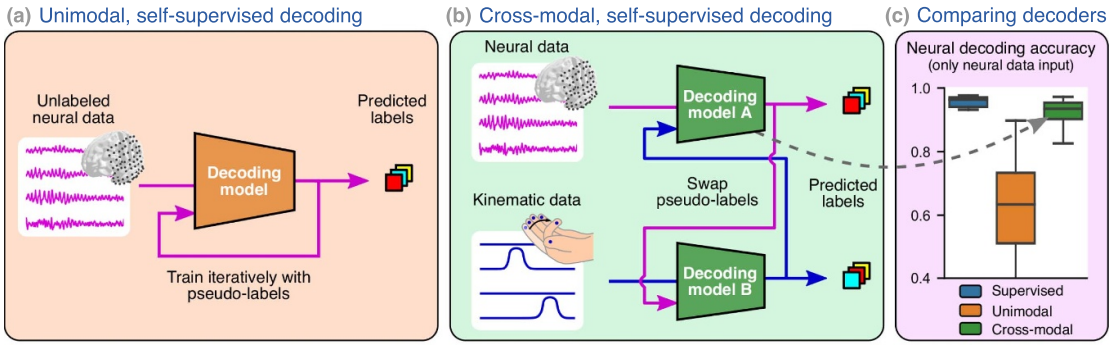

Figure 1. By sharing information across multiple data streams, we can train high-performing neural decoders from unlabeled data. (a) Self-supervised deep clustering iteratively generates pseudo-labels based on the current decoding model's outputs in order to train a neural decoder without the need for labeled data. For this study, we use HTNet [25] as the decoding model to predict movement-related behaviors from segmented neural data. (b) To improve model robustness, we applied cross-modal deep clustering [26], which swaps neural decoder pseudo-labels with those from additional kinematic and/or physiological data streams. (c) We find that neural decoders trained using cross-modal, self-supervised learning approach the performance of supervised neural decoders trained on labeled data. Note that when evaluating cross-modal neural decoders, we focus solely on the neural decoding model (decoding model A in (b)) and only feed in neural data; no pseudo-labels or other data streams are used.

approach provides a compelling alternative to tedious data annotation of neural recordings, enabling scalable analyses of large, complex neural datasets and robust brain–computer interfaces that can readily adapt to new real-world scenarios.

2. Results

Self-supervised learning with deep clustering uses patterns that emerge from the data to iteratively train a neural decoder, and cross-modal deep clustering takes further advantage of correlated patterns among multiple data streams. To understand this training approach, let us first consider a single stream of data, where a unimodal deep clustering decoding model is trained alongside a clustering algorithm that assigns each sample to a cluster based on the decoding model's output [38, 75]. Training proceeds iteratively, alternating between (a) optimizing the decoding model with the current cluster assignments (pseudo-labels) using backpropagation and (b) updating the pseudo-labels given the current decoding model [75, 76]. Pseudo-labels are constrained to equally partition the data to avoid the case where all events are assigned to one cluster. Similarly, cross-modal deep clustering uses pseudo-labels, but each decoding model is optimized using pseudo-labels from another data stream. In this way, after many iterations, this swapping of pseudo-labels directly ties together the data streams and the output of their decoding models [26]. Thus, cross-modal deep clustering provides a straightforward procedure to share information among multiple data streams while maintaining separate decoding models for each data stream.

We assessed cross-modal, self-supervised decoding performance on four datasets (table 1) and demonstrate in each case that cross-modal decoding outperforms unimodal, self-supervised models and approaches the accuracy of supervised models. We consider three movement decoding tasks: determining whether a participant's arm was moving or at rest (*ECoG move/rest* [56, 77] and *EEG move/rest* [73]), predicting which of five fingers was being flexed (*ECoG finger flexion* [40, 74]), and determining whether a participant was exposed to a visual or physical balance perturbation while either walking or standing (*EEG balance perturbations* [65]). Our decoding models all use the HTNet architecture, a compact convolutional neural network that has been demonstrated to perform well at decoding ECoG/EEG data [25, 78]. We consider ECoG move/rest, EEG move/rest, and ECoG finger flexion as proof-of-principle examples because the non-neural data is well-separated when clustered on its own; these are followed by a more challenging example that showcases the strengths of cross-modal clustering (EEG balance perturbations), especially when we extend the approach to more than two data streams.

We validated trained decoding models using accuracy on a withheld test set, which reflects how well each model generalizes to unseen data. We also assess each decoder's ability to effectively cluster unseen data using the v-measure [79]. To compute test accuracy for deep clustering models, we linearly mapped model output clusters to true labels from the training data and used this mapping to generate predictions with the test set [80]. During model training, we found that cross-modal, self-supervised decoders often converged to nearly identical accuracies across data streams for each decoding task (table S1). The differences in test accuracy that we report in this section primarily reflect how well each trained model is able to generalize to unseen data from its respective data stream.

Table 1. Multi-modal datasets used to test cross-modal neural decoding. We assessed the performance of supervised and self-supervised models using four multi-modal datasets during various movement tasks. Data streams include electrocorticography (ECoG), electroencephalography (EEG), electromyography (EMG), and multiple kinematic measurements. Kinematic measurements were obtained from markerless motion capture applied to video recordings (ECoG move/rest [56]), exoskeleton positions (EEG move/rest [73]), dataglove recordings (ECoG finger flexion [74]), and motion capture markers (EEG balance perturbations [65]). We computed the number of events per participant after balancing events across classes.

Dataset name	Task	# of classes	# of participants	# of data streams	Data streams	Events per participant (mean ± SD)
ECoG move/rest	Naturalistic arm movement or rest	2	12	2	ECoG, 2D arm position	1155 ± 568
EEG move/rest	Cued elbow flexion or rest	2	15	2	EEG, 3D arm position	118 ± 1
ECoG finger flexion	Cued flexion of individual fingers	5	8	2	ECoG, finger joint angles	114 ± 36
EEG balance perturbations	Rotations/pulls while standing/ walking	4	30	3	EEG, EMG, 3D body position	571 ± 89

2.1. Cross-modal decoding with two data streams

For all decoding tasks, we find that sharing self-supervised pseudo-labels among two data streams substantially improves decoding performance compared to unimodal, self-supervised models, while also approaching test accuracies of supervised models. In addition, we observe similar differences across model types for clustering performance (table S2).

2.1.1. Decoding arm movement vs. rest

In both move/rest decoding tasks, cross-modal, self-supervised neural decoders consistently outperform unimodal, self-supervised models and approach supervised model test accuracy (figure 2). For the naturalistic ECoG move/rest dataset, we find that neural decoder test accuracy is significantly affected by model type ($p = 3.25 \times 10^{-5}$; Friedman test [81]). Cross-modal ECoG decoders achieve an average accuracy of $88 \pm 9\%$ (median ± median absolute deviation (MAD)), which is well above random chance (50%). We find that cross-modal decoders have a small but significant decrease in test accuracy compared to supervised decoding performance ($91 \pm 6\%$, $p = 0.008$; Wilcoxon signed-rank test with false discovery rate correction [82, 83]). In contrast, the average test accuracy for unimodal, self-supervised decoders is only $57 \pm 5\%$, which is significantly lower than both cross-modal and supervised model performances ($p = 0.003$ for both). We find similar differences among models for the EEG move/rest dataset, with EEG test accuracy significantly affected by model type ($p = 1.73 \times 10^{-6}$). Again, the test accuracies of cross-modal ($76 \pm 9\%$) and supervised ($79 \pm 9\%$) decoders are well above random chance (50%), but do significantly differ from each other ($p = 0.005$). Still, both cross-modal and supervised decoders substantially outperform unimodal, self-supervised models ($49 \pm 6\%$, $p = 9.82 \times 10^{-4}$ for both comparisons), demonstrating the usefulness of cross-modal training for self-supervised neural decoding.

For both move/rest tasks, all three model types decode unseen arm position with over 90% accuracy. We find that this pose decoding performance is significantly affected by model type for EEG move/rest ($p = 9.65 \times 10^{-6}$), but not for the ECoG move/rest dataset ($p = 0.067$). For ECoG move/rest, move/rest decoding performance from two-dimensional (2D) arm position was $100 \pm 0\%$, $99 \pm 2\%$, and $98 \pm 3\%$ for supervised, unimodal, and cross-modal models, respectively. When decoding three-dimensional (3D) arm position for the EEG move/rest dataset, we find that cross-modal models ($94 \pm 5\%$) have a small but significant decrease in decoding performance compared to supervised ($99 \pm 1\%$, $p = 0.001$) and unimodal ($99 \pm 1\%$, $p = 0.001$) models. These high test accuracies for all decoders indicate how separable the move and rest classes are when decoding arm position. Cross-modal models are able to leverage this high separability in the arm position data stream to improve self-supervised neural decoding performance, which explains why cross-modal neural decoders notably outperform unimodal, self-supervised models.

2.1.2. Decoding finger flexion

When testing decoding performance on a more complex task, we find that pairwise cross-modal training again improves neural decoding performance substantially compared with unimodal, self-supervised models and nears supervised model performance (figure 3). We observe a significant effect of model type on neural decoding for this task ($p = 0.002$). Cross-modal neural decoders ($54 \pm 15\%$; median ± MAD) achieve substantially higher test accuracies than unimodal, self-supervised models ($24 \pm 4\%$) and approach supervised model performance ($60 \pm 8\%$, $p = 0.176$). Unimodal neural decoders perform near random chance (20%) for most participants, which is significantly worse than both cross-modal and supervised neural decoders ($p = 0.018$ for both).

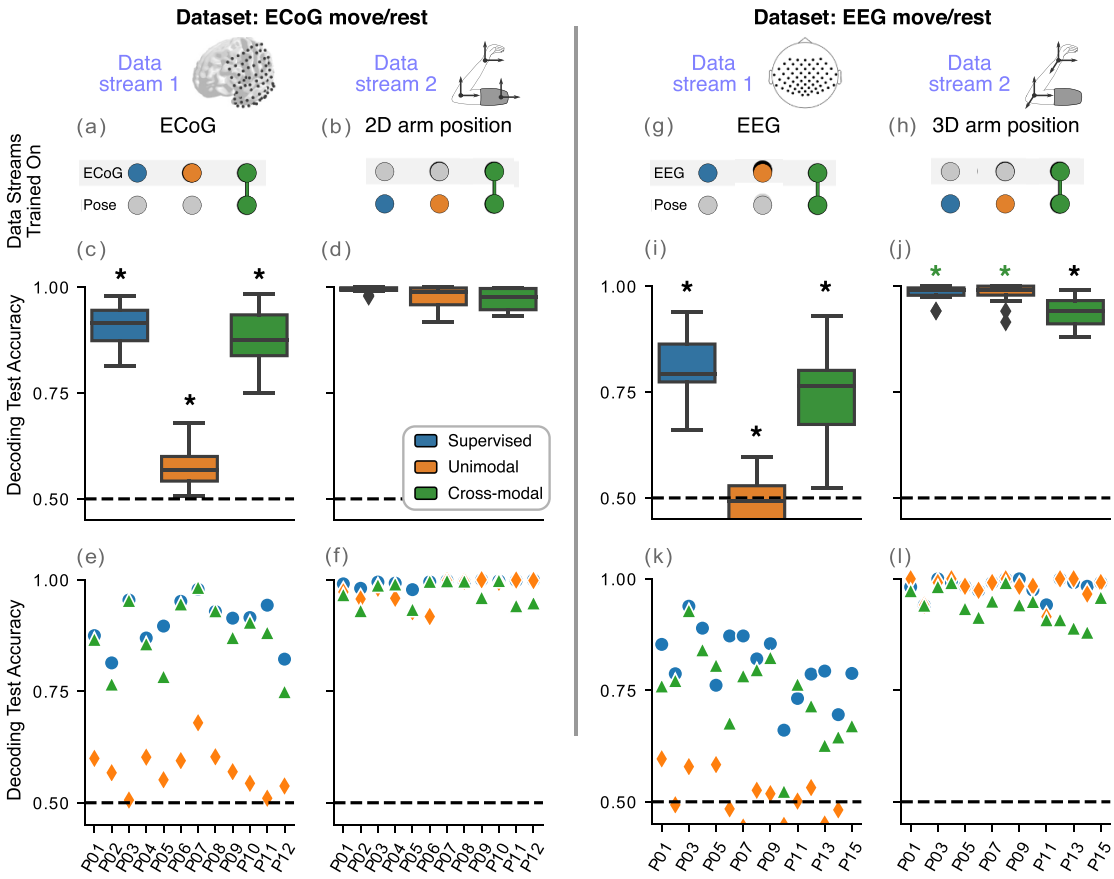

Figure 2. Cross-modal decoding of arm movement vs. rest approaches supervised decoding accuracies. Decoding models were trained on one or both of the two data streams, as depicted in (a), (b) for ECoG move/rest and (g), (h) for EEG move/rest. Decoding test accuracy is shown for ECoG move/rest (c), (d) and EEG move/rest (i), (j) datasets, averaged over ten random folds. Note that cross-modal training accuracy is quite similar between the two decoders from each data stream; the differences in test accuracy primarily reflect how well each decoder is able to generalize to unseen data from its respective data stream. For both datasets, the cross-modal decoders are able to leverage the high separability in the arm position data stream to achieve test accuracies well above unimodal, self-supervised models that approach supervised model performance. Asterisks denote significant differences from either specific models (shaded by color) or all other models (black). Single-participant performance is shown in (e), (f) and (k), (l) for the ECoG move/rest and EEG move/rest datasets, respectively. Random chance performance is denoted by the dashed, horizontal line at 50%.

Similar to decoding arm position for the move/rest datasets, all three model types can decode unseen finger joint angles at or above 90% accuracy. We observe a significant effect of model type on test accuracy when decoding joint angles ($p = 0.008$). Cross-modal joint angle decoders ($94 \pm 6\%$) have significant worse performance compared to supervised models ($100 \pm 0\%$, $p = 0.035$). Unimodal, self-supervised model accuracy ($95 \pm 7\%$) is similar to cross-modal performance ($p = 0.674$), indicating that supervised training yields more robust joint angle decoders than self-supervised learning. One explanation is this dataset contains a low number of events per class (23 ± 7 events per class) and that adding more training data would close the performance gap between supervised and self-supervised models. Regardless, cross-modal training is again able to leverage the high separability within the finger joint angle data stream to train high-quality neural decoders.

2.2. Decoding balance perturbations

Compared to the other datasets tested so far, the EEG balance perturbations dataset provides a much more challenging problem for cross-modal clustering because no data stream has near 100% accuracy when clustered on its own. This means that cross-modal decoders are relying on the shared information across data streams during training to boost decoding performance.

2.2.1. Cross-modal decoding with two data streams

Our results for decoding balance perturbations are consistent in observing that the pairwise cross-modal neural decoders outperform unimodal, self-supervised models and approach supervised model performance (figure 4). Because many participants performed this task, almost all comparisons of decoder test accuracy are statistically significant ($p < 0.05$). Still, we find that the pairwise cross-modal neural decoders (EEG/body pose:

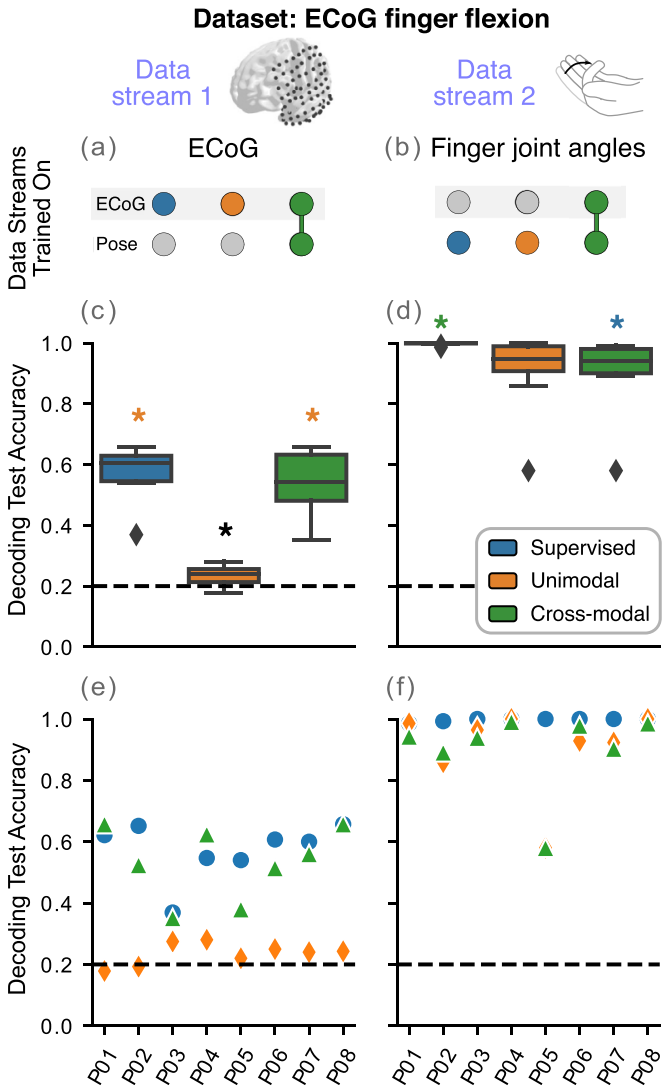

Figure 3. Cross-modal neural decoders achieve comparable performance to supervised models for decoding finger flexion. (a), (b) Decoding models were trained on one or both data streams. (c), (d) Cross-modal decoders again leverage the high separability in finger joint angles for different finger movements to nearly match supervised neural decoding performance. Asterisks denote significant differences from either specific models (shaded by color) or all other models (black). (e), (f) Decoding performance is shown for each participant, averaged over ten random folds. Random chance is indicated by the dashed, horizontal line at 20%.

$85 \pm 6\%$; EEG/electromyography (EMG): $83 \pm 9\%$) perform in between unimodal, self-supervised models ($63 \pm 16\%$) and supervised models ($98 \pm 2\%$). We also observe poor decoding performance near random chance (25%) for P02, P04, P11, and P15. These and other neural decoding outliers reflect the variable signal quality across participants during mobile EEG recordings, especially with only minimal preprocessing to reduce noise. Even so, pairwise cross-modal neural decoders using either body position or EMG clearly outperform unimodal, self-supervised models and demonstrate that any task-relevant data stream can be useful in improving self-supervised decoding performance.

Unlike the other datasets, pairwise cross-modal neural decoders approach supervised model performance despite using a non-neural data stream that is not easily separable across classes. Supervised decoding performance is well above chance (body position: $80 \pm 8\%$; EMG: $84 \pm 5\%$), but not at the near-perfect test accuracies seen for the other datasets. Similarly, unimodal, self-supervised models only achieve $\sim$50% test accuracy (body position: $52 \pm 9\%$; EMG: $45 \pm 5\%$), which is far below pairwise cross-modal accuracies for decoding body position (EEG/body pose: $77 \pm 10\%$; body pose/EMG: $70 \pm 11\%$) and EMG (EEG/EMG: $78 \pm 6\%$; body pose/EMG: $73 \pm 10\%$). These results demonstrate how cross-modal training can still develop high-quality self-supervised decoders, even without an easily separable data stream.

2.2.2. Cross-modal decoding with three data streams

By combining all three data streams from the EEG balance perturbations dataset, we improve cross-modal decoding performance to nearly match or even slightly exceed supervised model performance (figure 4). Again, almost all comparisons are

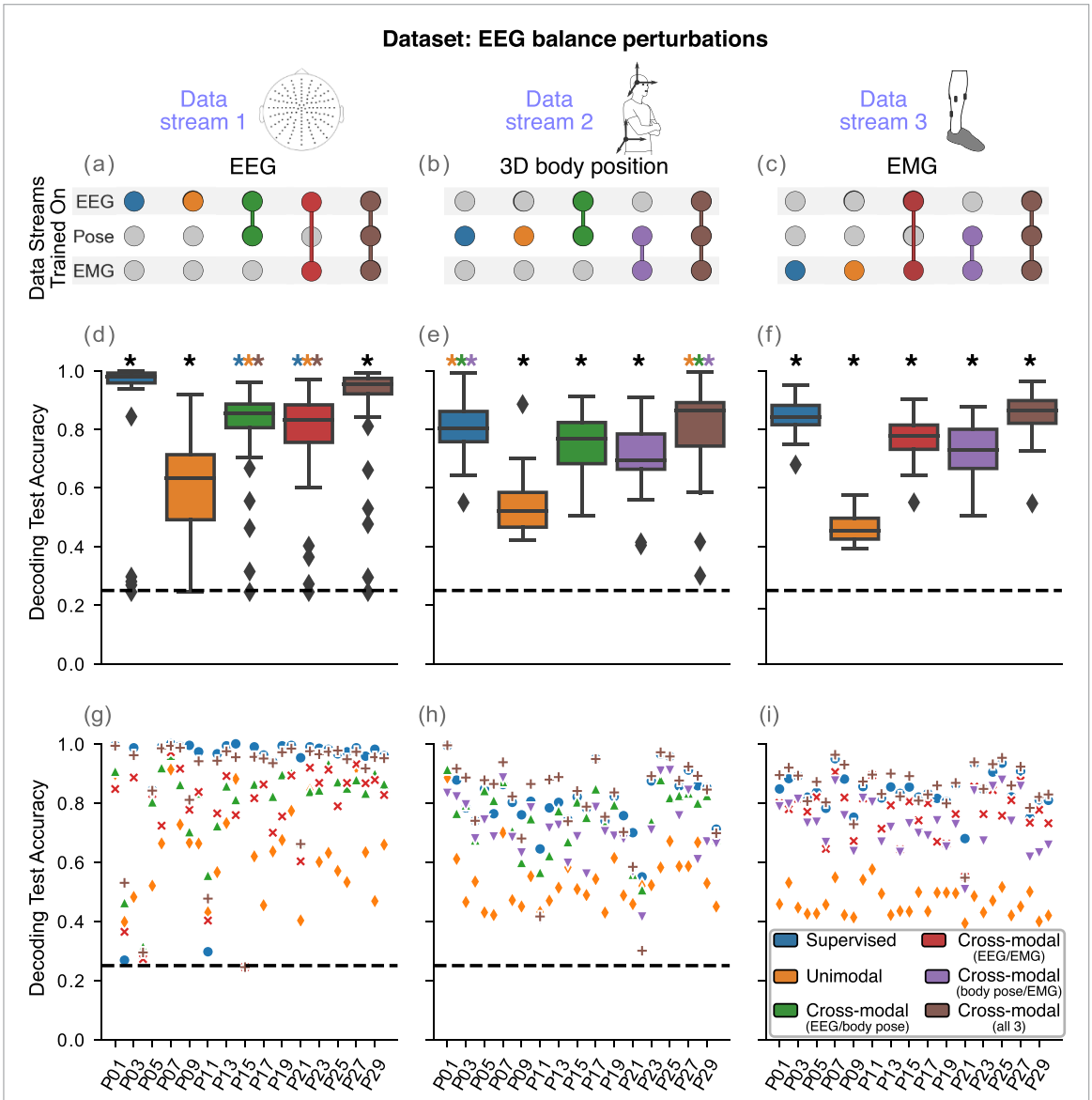

Figure 4. Cross-modal decoding across three data streams matches or even exceeds supervised model performance. (a)–(c) Decoding models were trained with different combinations of the three data streams (EEG, 3D body position, and EMG). (d)–(f) We performed cross-modal decoding using two or all three data streams. Cross-modal models using all three data streams perform similar to or slightly better than supervised decoders, demonstrating that sharing information among multiple data streams can lead to robust neural decoders even without labeled training data. Asterisks denote significant differences from either specific models (shaded by color) or all other models (black). (g)–(i) Single-participant test accuracies show the high variability decoding performance, especially for the EEG data stream. This high variability in EEG decoding performance is not surprising because recordings took place during perturbed balance-beam walking and are likely contaminated with noise from head motion, neck muscles, and eye movements. Random chance is denoted by the dashed, horizontal line at 25%.

statistically significant ($p < 0.05$) due to how many participants are in this dataset. For decoding EEG, trimodal neural decoding performance ($95 \pm 3\%$) is substantially higher than unimodal and pairwise cross-modal accuracies. We do find a small but significant decrease in trimodal performance compared with supervised decoders ($p = 0.003$), but both decoders are within ~1–3% for most participants. In addition, trimodal decoding performance for the non-neural data streams is significantly increased compared to pairwise cross-modal decoders (body position: $86 \pm 10\%$, $p = 2.20 \times 10^{-5}$; EMG: $86 \pm 6\%$, $p = 1.40 \times 10^{-5}$). For decoding body position, trimodal decoding performance is not significantly different compared to supervised decoder accuracy ($p = 0.131$). For EMG decoding, we actually observe a small but significant increase in trimodal decoding accuracy over supervised decoding performance ($p = 0.004$). We also find similar differences among model types for clustering performance (table S3). Taken together, our findings demonstrate that including additional data streams can lead to robust self-supervised decoders that do not require labeled training data.

2.3. Effect of expected number of clusters on performance

We also assess how the expected number of clusters (K) impacts pairwise cross-modal performance, finding that selecting a value for K that is above the true

number of classes notably affects test accuracy but not clustering performance (figure S1). We performed this assessment on the ECoG finger flexion dataset. For both data streams, setting K to less than the number of classes led to decreases in test accuracy and clustering performance relative to cross-modal models with K equal to the number of classes. In addition, we found that over-clustering, or setting K to greater than the number of classes, also decreases test accuracy compared with K equal to the number of classes, likely because the model had difficulty generalizing clusters that divided up the same class. In contrast, clustering performance for over-clustered models increases or stays the same compared to models with K equal to the number of classes. These findings highlight the importance of carefully selecting K depending on the overall goal; if generalized decoding is desired, then K should be carefully chosen, but if clustering is the primary objective, then choosing a large K should be sufficient.

3. Discussion

In this paper, we demonstrate that cross-modal deep clustering extends well to neural decoding applications. We show four examples where cross-modal neural decoders achieve performance that approaches supervised models despite using no labeled data. Indeed, cross-modal clustering works well when one of the data streams can be clustered well on its own (figures 2 and 3) as well as when none are (figure 4). Because neural recordings are routinely measured simultaneously with multiple other data streams, we also extend this cross-modal approach to include more than two data streams and find notably improved performance when adding a third data stream to train cross-modal neural decoders. It is also worth noting that cross-modal decoding still performs well when both data streams are high-dimensional and do not separate well during unimodal decoding (e.g. figure 4 EEG-EMG decoder). Our findings demonstrate that including additional, task-relevant data streams during cross-modal training leads to robust, self-supervised neural decoders that do not require labeled data.

To our knowledge, cross-modal deep clustering is the first self-supervised approach that can create high-performing neural decoders from unlabeled data without relying on contrastive learning and data augmentations. This approach seems to perform well because variations among different, task-relevant data streams reduce overfitting during model training, similar to what is achieved by data augmentation [52, 53]. Even supervised neural decoders can overfit to the inherent variability of neural data, as well as task-irrelevant changes in neural activity due to factors such as fatigue and emotional state [14, 84, 85]. These task-irrelevant patterns are less likely to appear across multiple data streams, thus reducing the tendency to overfit. This hypothesis also explains why adding more data streams improves cross-modal decoder performance for the EEG balance perturbations dataset (figure 4) and underlies many sensor fusion and multi-modal approaches for reliable neural decoding [70–72, 86, 87]. Our proposed approach for cross-modal deep clustering using any number of data streams (figure 5) leverages the multiple neural, physiological, and kinematic data streams often collected in neuroscience research studies [24, 64, 66–69, 88, 89].

Cross-modal deep clustering can also be adapted to a wide variety of different model architectures and data streams. Here, we used a convolutional neural network (HTNet) for decoding [25], but we could have applied a larger convolutional neural network model or added recurrent layers [90, 91]. Such adjustments could potentially improve decoding accuracy, but would increase training time and possibly require more training data to train. HTNet is also able to pool neural data across multiple participants, raising the possibility of training neural decoders on unlabeled, multi-participant data [25]. In addition to its flexibility in model architecture, cross-modal deep clustering can also be applied to any data streams that are synchronously recorded. We chose kinematic data streams such as body position and joint angle because of their high relevance to movement behaviors [92]. Ideal data streams contain high-quality, task-relevant information. For exploratory analyses of large neural datasets, selecting certain data streams could also be used to bias models towards decoding specific behaviors.

Despite its impressive decoding performance on unlabeled data, cross-modal deep clustering has multiple limitations to consider. First, like with many clustering algorithms, the number of expected clusters must be chosen with extreme care to obtain high-quality decoding performance on unseen data (figure S1). One way to decide on a reasonable number of expected clusters without data labels is to generate clusters for a subset of examples, train a classifier using the cluster assignments, and test how well this classifier's predictions match the cluster assignments for the unseen data. This cross-validation approach has been used for selecting the optimal number of clusters during k-means clustering [93]. Another limiting factor of our current approach is that it requires segmented, event-related neural data that transitions from a consistent baseline rest state. We could have potentially extended our approach to continuous data using a sliding window [94, 95], but we would have to account for event transitions from non-rest states [96]. Finally, our approach may not be appropriate for decoding imagined movements, where kinematic data streams are not available. Still, studies have shown that other data streams such as eye gaze,

heart rate, and muscle activity can be affected by imagined movements [97–99] and thus may be useful for cross-modal decoding of imagined movements.

This cross-modal deep clustering approach can be extended to a wide variety of neural decoding applications where limited labeled data is available, including predicting mood and somatosensation [100, 101]. In addition, fine-tuning and other transfer learning techniques can be readily applied to improve trained model performance, as is often done for self-supervised models [52, 53]. Our current approach of randomly sharing pseudo-labels for more than two data streams may be made more rigorous by weighting each data stream based on its reliability and data quality. Data augmentations could also be useful in improving decoder performance. While we focused on self-supervised learning without data augmentations, deep clustering has been frequently applied to augmented data [38, 39, 75]. Using multiple data streams should help minimize any biases introduced by a specific data augmentation. Overall, cross-modal deep clustering is a promising alternative to tedious data annotation and provides a reliable framework for achieving high-quality neural decoding when curating an annotated dataset is difficult or infeasible.

4. Methods

4.1. Cross-modal deep clustering

We implemented the unimodal deep clustering approach from Asano *et al* [75] (figure 5(b)). This unimodal approach alternates between (a) optimizing the decoding model given the current pseudo-labels and (b) updating the pseudo-labels given the current decoding model. This first step involves minimizing cross-entropy loss, similar to supervised training (figure 5(a)). In the second step, the Sinkhorn–Knopp algorithm takes in the model outputs and updates the label assignments with the constraint that the pseudo-labels must equally partition the data among the expected number of clusters [75, 76]. This constraint avoids situations where all events are assigned the same pseudo-label and the model minimizes cross-entropy loss by always predicting the same label [75]. We primarily set the number of expected clusters (K) equal to the number of true clusters and later explored the effects of K on model performance. All models were created and trained using Python 3.8.5 and PyTorch 1.7.1.

We modified this unimodal approach from Asano *et al* [75] to share pseudo-labels across multiple modalities as done in Alwassel *et al* [26] (figure 5(c)). This cross-modal deep clustering approach was found to outperform other multi-modal deep clustering models [26]. Cross-modal deep clustering involves swapping pseudo-labels between the two data streams when optimizing the decoding model. This swapping directly affects not only how the decoding model is learned, but also indirectly what the pseudo-labels are because they are generated from the decoding model's output.

We extend this pairwise cross-modal deep clustering approach to accommodate more than two data streams. For each event, we randomly select a pseudo-label to train the decoding model for a particular data stream from any of the other data streams. This pseudo-label selection process randomly switches among the other N-1 data streams when selecting pseudo-labels to train the current decoding model, combining information from multiple data streams during training. When there are only two data streams, the two models simply swap pseudo-labels, just like the previous pairwise approach [26]. With our extended N-wise cross-modal deep clustering approach, we assume that the decoding models for every data stream will converge to similar pseudo-labels by the end of training. Otherwise, the trained models may be biased to the most recent pseudo-labels used during training.

4.2. HTNet decoding model

For the decoding model, we used HTNet [25], a convolutional neural network for generalized neural decoding (figure S2). We chose HTNet because it has few parameters and identifies data-driven spatiotemporal features in the frequency domain, where neural data is often most easily separated. We also applied HTNet for decoding all non-neural data streams to maintain consistency across datasets. We used the user-defined number of expected clusters to determine the size of HTNet's output layer. That way, HTNet generates one value for each expected cluster, which are transformed using the softmax function into joint probabilities estimated by the model. These estimated probabilities are compared to the pseudo-label for event and cross-entropy loss between the two is minimized [75].

We slightly modified the original HTNet architecture by replacing the initial temporal convolution layer with SincNet filters [102]. Instead of learning the entire convolution kernel, each SincNet filter learns only the minimum and maximum frequency cutoffs for a pre-defined band-pass filter. Compared with a temporal convolution layer, SincNet filters are less biased towards low frequencies because they do not need to learn the complex kernel shape often needed for isolating high frequencies. We found that SincNet filters improve HTNet's performance, especially for data streams such as ECoG with task-relevant features at high-frequencies.

4.3. Datasets

We assess cross-modal deep clustering performance on four datasets that span a variety of tasks and recording modalities (table 1). All four datasets are publicly available and include at least one other synchronized data stream besides the neural recording. Note that we did not discard noisy events for any

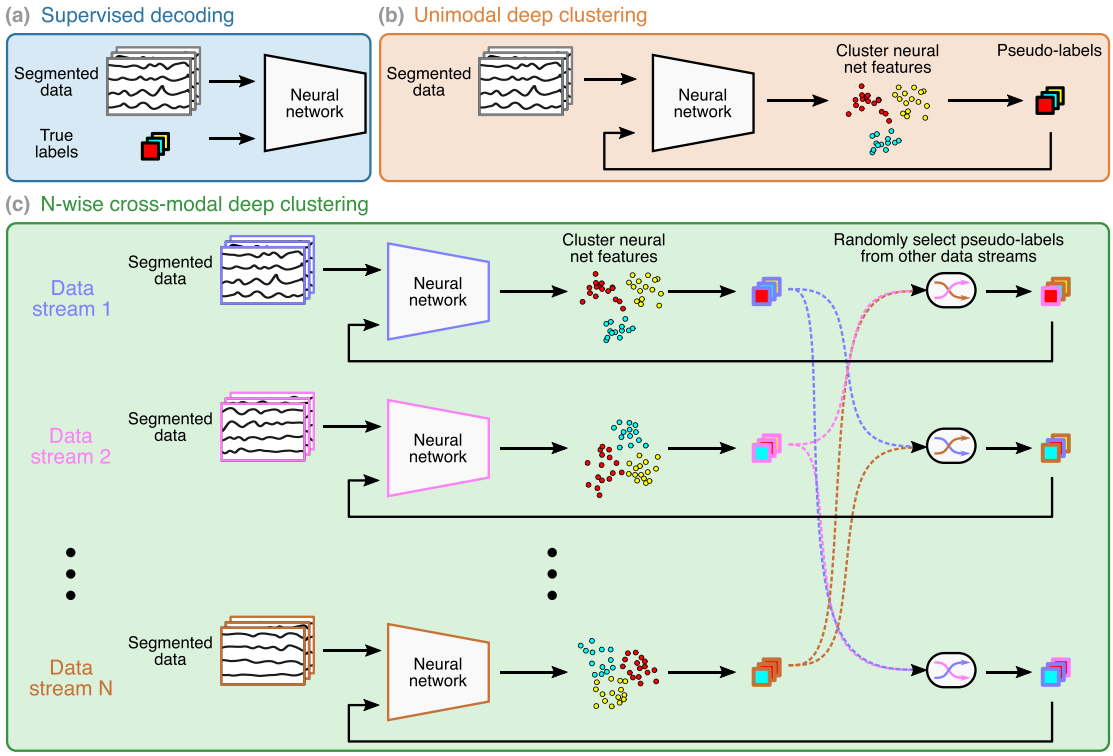

Figure 5. Supervised and self-supervised training paradigms. We compare performance of supervised neural decoders with unimodal and cross-modal self-supervised deep clustering models. (a) Supervised decoders are trained using the true data labels. (b) In contrast, deep clustering self-generates pseudo-labels by clustering the neural network model outputs and then uses these pseudo-labels to train the model. Note that the neural network and clustering technique are alternately optimized during training. (c) In cross-modal deep clustering, decoder models for each data steam are trained using pseudo-labels randomly selected from all other data streams except itself. For all supervised and self-supervised approaches, we used HTNet, a convolutional neural network for decoding ECoG and EEG data, as the neural network decoder.

dataset. Variability in the number of events across participants reflects either differences in the number of events actually recorded or events that were discarded because not enough data was available to segment.

4.3.1. ECoG move vs. rest dataset

Concurrent intracranial ECoG and upper-body positions were obtained from 12 human participants (8 males, 4 females; aged 29 ± 8 years old (mean $\pm$ SD)) during continuous clinical epilepsy monitoring conducted at Harborview Medical Center in Seattle, WA [56, 77, 103]. ECoG electrodes were implanted primarily in one hemisphere (5 right, 7 left) and included cortical surface recordings along with penetrating electrodes into subcortical areas which are also known as stereoelectroencephalography. This study was approved by the University of Washington Institutional Review Board, and all participants provided written informed consent. ECoG and pose data were sampled at 500 Hz and 30 FPS, respectively. This dataset is publicly available at: https://doi.org/10.6084/m9.figshare.16599782.

The goal is to discriminate between rest events and movements of the arm contralateral to the implanted electrode hemisphere. Move/rest classifications were determined via markerless pose tracking of video recordings and automated state segmentation [77]. Move events were defined as wrist movements that followed 0.5 s of no movement. In contrast, rest events denoted periods of 3 or more seconds with no movement in either wrist. We balanced the number of move/rest events using random undersampling [104], leading to 1155 ± 568 events per participant.

ECoG data was preprocessed by removing DC drift, high-amplitude discontinuities, band-pass filtering (1–200 Hz), notch filtering, referencing to the common median across electrodes, and downsampling to 250 Hz [56]. Note that this downsampling procedure and all other resampling done in this study included an anti-aliasing filter that removed high-frequency content above the new Nyquist rate before downsampling the signal. We selected 2D pose trajectories obtained from video recordings for three keypoints (shoulder, elbow, and wrist of the contralateral arm) and computed the difference between neighboring timepoints. Data from both modalities were trimmed to 2 s segments centered around each event.

4.3.2. EEG move vs. rest dataset

Concurrent EEG and arm positions were recorded from 15 human participants during visually cued

elbow flexion (6 males, 9 females; aged 27 ± 5 years old (mean ± SD)). EEG data was recorded from 61 electrodes and sampled at 512 Hz. For pose data, we selected 3D elbow and wrist positions recorded by an exoskeleton during the experiment and sampled at 512 Hz. This study was approved by the ethics committee of the Medical University of Graz, and all participants provided written informed consent [73]. This dataset is publicly available at: http://bnci-horizon-2020.eu/database/data-sets.

Similar to the previous dataset, the decoding task involves two-class classification of either move or rest events. Here, a move event corresponds to cued elbow flexion of the right arm. Each participant performed 120 total trials (60 movement and 60 rest trials). We aligned data segments to movement onset, which was identified by thresholding the wrist's radial displacement after the visual cue to move.

EEG data was notch filtered at 50 Hz, referenced to the right mastoid, and band-pass filtered (0.01–200 Hz). We performed further processing by high-pass filtering at 1 Hz, average referencing, and resampling to 250 Hz. 3D pose trajectories were obtained from each participant's right elbow and hand, and we computed the difference between neighboring timepoints to quantify the change in position.

4.3.3. ECoG finger flexion dataset

ECoG and finger joint angles were recorded concurrently from nine human participants during visually cued finger flexion (three males, six females; aged 27 ± 9 years old). ECoG electrodes were implanted primarily in one hemisphere (2 right, 7 left) and included only cortical surface recordings (38–64 electrodes per participant). Finger joint angles were recorded using a 5 degree-of-freedom dataglove sensor [74]. All patients participated in a purely voluntary manner, after providing informed written consent, under experimental protocols approved by the Institutional Review Board of the University of Washington (#12193). All patient data was anonymized according to IRB protocol, in accordance with HIPAA mandate. These data originally appeared in the manuscript 'Human Motor Cortical Activity Is Selectively Phase-Entrained on Underlying Rhythms' published in PLoS Computational Biology in 2012 [74]. Both ECoG and finger joint angles were originally sampled at 1000 Hz. This dataset is publicly available at: https://searchworks.stanford.edu/view/zk881ps0522 [40].

The decoding task here is to classify which of five fingers is being flexed. All finger flexions occurred in the hand contralateral to the ECoG implantation hemisphere. Each cued finger flexion lasted 2 s, followed by 2 s of rest. Every participant performed 150 pseudo-randomly interleaved finger flexions (30 for each finger).

We discarded one participant (mv) due to an inability to obtain cued movement times. ECoG data was band-pass filtered between 4–250 Hz, notch filtered at 60 Hz and its harmonics, average referenced, and downsampled to 250 Hz. Pose data was standardized and also downsampled to 250 Hz. We removed finger flexion events where no movement occurred.

4.3.4. EEG balance perturbations dataset

Concurrent EEG, EMG, and 3D body position were collected from 30 human participants during sensorimotor perturbations to standing and walking balance (15 males, 15 females; aged 23 ± 5 years old (mean ± SD)) [65]. EEG data was recorded from 128 electrodes and sampled at 512 Hz. EMG was sampled at 1 kHz and collected from four muscles on each leg (8 total): tibialis anterior, soleus, medial gastrocnemius, and peroneus longus. 3D body position of the head and sacrum was recorded using motion capture markers with 100 Hz sampling. This protocol was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board, and all subjects provided written informed consent [65]. This dataset is publicly available at: https://openneuro.org/datasets/ds003739 [105].

The decoding task for this dataset is to classify between two sensorimotor perturbations during either standing or walking (four total classes). Sensorimotor perturbations involved either a 1 s mediolateral pull at the waist or a half-second 20 degree field-of-view rotation using a virtual reality headset. Each class includes 150 events recorded during a separate 10 min recording session and balanced between left/right pulls or clockwise/counterclockwise rotations.

EEG data were downsampled to 256 Hz, 1 Hz high-pass filtered, and referenced to the common median across electrodes. We removed the linear trends in EMG and pose data using Matlab 2013a. EMG and pose data were then resampled to match the EEG sampling rate of 256 Hz.

4.4. Model training and comparisons

We compared decoder model performance for three types of training: supervised, unimodal self-supervised, and cross-modal self-supervised. All decoder hyperparameters were fixed across these three models; the only difference was using either the true labels or self-generated pseudo-labels during model training. For training, supervised and unimodal, self-supervised models were trained for 40 epochs, while cross-modal models were trained for 200 epochs. We did not use any stopping criteria and selected 40 epochs based on the average number of epochs used to train HTNet on similar neural data where the likelihood of overfitting was minimized by applying early stopping based on improvement in

validation set accuracy [25]. For cross-modal models, we performed model training over many more epochs in order to provide enough time to converge across all data streams. Unlike unimodal and supervised models, we found that cross-modal models did not overfit when trained over hundreds of epochs. For every dataset, we split the data into ten stratified folds to minimize bias in our selection of training/testing data (90% of events for training, 10% for testing).

4.5. Model hyperparameters

We set decoder model (HTNet) hyperparameters for the two move/rest datasets based on a previous decoding study [25]. See table S4 for more detailed information about the hyperparameters used. For the ECoG finger flexion encoder, we increased the number of spatial filters from 2 to 5 to adequately capture different spatial features across finger movements. To keep the total number of fitted parameters low, we decreased the number of temporal filters to 6. In addition, we did not use SincNet and spectral power for the arm position data to preserve the low-frequency features of the data. For the EEG balance dataset, we similarly increased the number of spatial filters to 3 and decreased the number of temporal filters to 6.

4.6. Model validation metrics

We chose two metrics for validating trained decoding model performance: test accuracy and v-measure. Both model validation metrics are always averaged over ten folds, selected with stratified random sampling to preserve the percentage of samples for each class within each folds. Test accuracy measures model classification accuracy on unseen data, while v-measure assesses the model's ability to separately cluster different classes. For accuracy, we computed the linear mapping between true and predicted clusters that yielded the maximum accuracy on the training set [80]. Then, we computed the accuracy on the test set using the same mapping between true and predicted clusters. For v-measure, we computed the harmonic mean of completeness (how much of each class is present in a single cluster) and homogeneity (how much each cluster is made up of a single class) [79]. V-measure can range between 0 and 1, with 1 indicating perfect clustering. We focus primarily on test accuracy instead of clustering performance because we want trained decoding models that can generalize to unseen data.

4.7. Statistical tests

We used non-parametric hypothesis tests to identify significant differences in performance between models. First, we first test if performance significantly changes across models at all using a Friedman test [81], which is the non-parametric equivalent to a repeated measures analysis of variance (ANOVA). If the Friedman test result was significant ($p < 0.05$), we then tested for pairwise differences between model types using a two-sided Wilcoxon signed-rank test, which is the non-parametric analogue to a pairwise t-test. We also controlled for spuriously significant p-values when performing multiple pairwise comparisons by applying a false discovery rate correction [82, 83].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.18112/openneuro.ds003739.v1.0.3.

Code and data availability

Our code is publicly available at: https://github.com/BruntonUWBio/cross-modal-ssl-htnet. The code in this repository can be used in conjunction with publicly available ECoG [40, 77] and EEG [73, 106] datasets to generate all of the main findings and figures from our study.

Acknowledgments

We thank Satpreet Singh, Zoe Steine-Hanson, Ellie Strandquist, and Pierre Karashchuk for discussions and help with data analysis and model design. This research was supported by funding from the National Science Foundation (1630178 and EEC-1028725), the Washington Research Foundation, and the Weill Neurohub.

ORCID iDs

Steven M Peterson https://orcid.org/0000-0003-0782-5788
Rajesh P N Rao https://orcid.org/0000-0003-0682-8952
Bingni W Brunton https://orcid.org/0000-0002-4831-3466

References

[1] Ganzer P D, Colachis S C, Schwemmer M A, Friedenberg D A, Dunlap C F, Swiftney C E, Jacobowitz A F, Weber D J, Bockbrader M A and Sharma G 2020 Restoring the sense of touch using a sensorimotor demultiplexing neural interface *Cell* **181** 763–73.e12
[2] Miller K J, Hermes D and Staff N P 2020 The current state of electrocorticography-based brain–computer interfaces *Neurosurg. Focus* **49** E2
[3] Volkova K, Lebedev M A, Kaplan A and Ossadtchi A 2019 Decoding movement from electrocorticographic activity: a review *Front. Neuroinform.* **13** 74
[4] Niketeghad S and Pouratian N 2019 Brain machine interfaces for vision restoration: the current state of cortical visual prosthetics *Neurotherapeutics* **16** 134–43
[5] Martin S, Millán J D R, Knight R T and Pasley B N 2019 The use of intracranial recordings to decode human language: challenges and opportunities *Brain Lang.* **193** 73–83
[6] Sani O G, Yang Y, Lee M B, Dawes H E, Chang E F and Shanechi M M 2018 Mood variations decoded from

multi-site intracranial human brain activity *Nat. Biotechnol.* **36** 954–61
[7] Moses D A *et al* 2021 Neuroprosthesis for decoding speech in a paralyzed person with anarthria *New Engl. J. Med.* **385** 217–27
[8] Degenhart A D, Bishop W E, Oby E R, Tyler-Kabara E C, Chase S M, Batista A P and Byron M Y 2020 Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity *Nat. Biomed. Eng.* **4** 672–85
[9] Oby E R, Hennig J A, Batista A P, Yu Byron M and Chase S M 2020 Intracortical brain–machine interfaces *Neural Engineering* (Berlin: Springer) pp 185–221
[10] Collinger J L, Gaunt R A and Schwartz A B 2018 Progress towards restoring upper limb movement and sensation through intracortical brain-computer interfaces *Curr. Opin. Biomed. Eng.* **8** 84–92
[11] Gu X, Cao Z, Jolfaei A, Xu P, Wu D, Jung T-P and Lin C-T 2020 EEG-based brain-computer interfaces (BCIs): a survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications (arXiv:2001.11337)
[12] Rao R P N 2013 *Brain-Computer Interfacing: An Introduction* (Cambridge: Cambridge University Press)
[13] Pandarinath C, Cora Ames K, Russo A A, Farshchian A, Miller L E, Dyer E L and Kao J C 2018 Latent factors and dynamics in motor cortex and their application to brain–machine interfaces *J. Neurosci.* **38** 9390–401
[14] Glaser J I, Benjamin A S, Chowdhury R H, Perich M G, Miller L E and Kording K P 2020 Machine learning for neural decoding *eNeuro* **7** ENEURO.0506-19.2020
[15] Bigdely-Shamlo N, Cockfield J, Makeig S, Rognon T, la Valle C, Miyakoshi M and Robbins K A 2016 Hierarchical event descriptors (HED): semi-structured tagging for real-world events in large-scale EEG *Front. Neuroinform.* **10** 42
[16] Karashchuk P, Rupp K L, Dickinson E S, Walling-Bell S, Sanders E, Azim E, Brunton B W and Tuthill J C 2021 Anipose: a toolkit for robust markerless 3D pose estimation *Cell Reports* **36** 109730
[17] Wu D, Xu Y, and Lu B-L 2020 Transfer learning for EEG-based brain-computer interfaces: a review of progress made since 2016 (arXiv:2004.06286)
[18] van Erp J, Lotte F and Tangermann M 2012 Brain-computer interfaces: beyond medical applications *Computer* **45** 26–34
[19] Huang G, Liu G, Meng J, Zhang D and Zhu X 2010 Model based generalization analysis of common spatial pattern in brain computer interfaces *Cogn. Neurodyn.* **4** 217–23
[20] Cohen M X 2014 *Analyzing Neural Time Series Data: Theory and Practice* (Cambridge, MA: MIT Press)
[21] Jing L and Tian Y 2021 Self-supervised visual feature learning with deep neural networks: a survey *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 4037–58
[22] Liu X, Zhang F, Hou Z, Mian Li, Wang Z, Zhang J and Tang J 2021 Self-supervised learning: generative or contrastive *IEEE Trans. Knowl. Data Eng.* **33** 1
[23] Lacourse K, Yetton B, Mednick S and Warby S C 2020 Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data *Sci. Data* **7** 1–14
[24] Wang N, Farhadi A, Rao R and Brunton B 2018 Ajile movement prediction: multimodal deep learning for natural human neural recordings and video *Proc. AAAI Conf. on Artificial Intelligence* vol 32
[25] Peterson S M, Steine-Hanson Z, Davis N, Rao R P N and Brunton B W 2021 Generalized neural decoders for transfer learning across participants and recording modalities *J. Neural Eng.* **18** 026014
[26] Alwassel H, Mahajan D, Korbar B, Torresani L, Ghanem B and Tran D 2020 Self-supervised learning by cross-modal audio-video clustering *Advances in Neural Information Processing Systems* vol 33
[27] Mikolov T, Sutskever I, Chen K, Corrado G S and Dean J 2013 Distributed representations of words and phrases and their compositionality *Advances in Neural Information Processing Systems* pp 3111–9
[28] Devlin J, Chang M-W, Lee K and Toutanova K 2018 BERT: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)
[29] Brown T B *et al* 2020 Language models are few-shot learners (arXiv:2005.14165)
[30] Kingma D P and Welling M 2013 Auto-encoding variational Bayes (arXiv:1312.6114)
[31] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* vol 27
[32] He K, Fan H, Wu Y, Xie S and Girshick R 2020 Momentum contrast for unsupervised visual representation learning *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 9729–38
[33] Chen T, Kornblith S, Norouzi M and Hinton G 2020 A simple framework for contrastive learning of visual representations *Int. Conf. on Machine Learning* (PMLR) pp 1597–607
[34] Chicco D 2021 Siamese neural networks: an overview *Artificial Neural Networks* (Berlin: Springer) pp 73–94
[35] Banville H, Chehab O, Hyvärinen A, Engemann D-A and Gramfort A 2021 Uncovering the structure of clinical EEG signals with self-supervised learning *J. Neural Eng.* **18** 046020
[36] Goyal P *et al* 2021 Self-supervised pretraining of visual features in the wild (arXiv:2103.01988)
[37] Grill J-B *et al* 2020 Bootstrap your own latent: a new approach to self-supervised learning (arXiv:2006.07733)
[38] Caron M, Bojanowski P, Joulin A and Douze M 2018 Deep clustering for unsupervised learning of visual features *Proc. European Conf. on Computer Vision (ECCV)* pp 132–49
[39] Caron M, Misra I, Mairal J, Goyal P, Bojanowski P and Joulin A 2020 Unsupervised learning of visual features by contrasting cluster assignments (arXiv:2006.09882)
[40] Miller K J 2019 A library of human electrocorticographic data and analyses *Nat. Hum. Behav.* **3** 1225–35
[41] Parvizi J and Kastner S 2018 Promises and limitations of human intracranial electroencephalography *Nat. Neurosci.* **21** 474–83
[42] Takaura K, Tsuchiya N and Fujii N 2016 Frequency-dependent spatiotemporal profiles of visual responses recorded with subdural ECoG electrodes in awake monkeys: differences between high- and low-frequency activity *NeuroImage* **124** 557–72
[43] Gunduz A, Brunner P, Daitch A, Leuthardt E C, Ritaccio A L, Pesaran B and Schalk G 2011 Neural correlates of visual-spatial attention in electrocorticographic signals in humans *Front. Hum. Neurosci.* **5** 89–89
[44] Pistohl T, Ball T, Schulze-Bonhage A, Aertsen A and Mehring C 2008 Prediction of arm movement trajectories from ECoG-recordings in humans *J. Neurosci. Methods* **167** 105–14
[45] Raghu S, Sriraam N, Gommer E D, Hilkman D M W, Temel Y, Rao S V, Hegde A S and Kubben P L 2020 Cross-database evaluation of EEG based epileptic seizures detection driven by adaptive median feature baseline correction *Clin. Neurophysiol.* **131** 1567–78
[46] Kaplan A Y, Fingelkurts A A, Fingelkurts A A, Borisov S V and Darkhovsky B S 2005 Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges *Signal Process.* **85** 2190–212
[47] Cole S and Voytek B 2019 Cycle-by-cycle analysis of neural oscillations *J. Neurophysiol.* **122** 849–61

[48] Donoghue T *et al* 2020 Parameterizing neural power spectra into periodic and aperiodic components *Nat. Neurosci.* **23** 1655–65
[49] Voytek B, Kramer M A, Case J, Lepage K Q, Tempesta Z R, Knight R T and Gazzaley A 2015 Age-related changes in 1/*f* neural electrophysiological noise *J. Neurosci.* **35** 13257–65
[50] Dubey A and Ray S 2019 Cortical electrocorticogram (ECoG) is a local signal *J. Neurosci.* **39** 4299–311
[51] Troy M L, Joseph T G and Daniel P F 2012 How many electrodes are really needed for EEG-based mobile brain imaging? *J. Behav. Brain Sci.* **2** 387–93
[52] Cheng J Y, Goh H, Dogrusoz K, Tuzel O and Azemi E 2020 Subject-aware contrastive learning for biosignals (arXiv:2007.04871)
[53] Kostas D, Aroca-Ouellette S and Rudzicz F 2021 BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data (arXiv:2101.12037)
[54] Mohsenvand M N, Izadi M R and Maes P 2020 Contrastive representation learning for electroencephalogram classification *Proc. Machine Learning for Health* (PMLR) pp 238–53
[55] Han J, Gu X and Lo B 2021 Semi-supervised contrastive learning for generalizable motor imagery EEG classification *2021 IEEE 17th Int. Conf. on Wearable and Implantable Body Sensor Networks (BSN)* (IEEE) pp 1–4
[56] Peterson S M, Singh S H, Wang N X R, Rao R P N and Brunton B W 2021 Behavioral and neural variability of naturalistic arm movements *eNeuro* **8** ENEURO.0007-21.2021
[57] Ratcliff R, Philiastides M G and Sajda P 2009 Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG *Proc. Natl Acad. Sci.* **106** 6539–44
[58] Reed C J, Metzger S, Srinivas A, Darrell T and Keutzer K 2021 SelfAugment: automatic augmentation policies for self-supervised learning *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 2674–83
[59] Araslanov N and Roth S 2021 Self-supervised augmentation consistency for adapting semantic segmentation *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 15384–94
[60] Lashgari E, Liang D and Maoz U 2020 Data augmentation for deep-learning-based electroencephalography *J. Neurosci. Methods* **346** 108885
[61] Liu R, Azabou M, Dabagia M, Lin C-H, Azar M G, Hengen K, Valko M and Dyer E 2021 Drop, swap and generate: a self-supervised approach for generating neural activity *Advances in Neural Information Processing Systems* vol 34
[62] Artoni F, Fanciullacci C, Bertolucci F, Panarese A, Makeig S, Micera S and Chisari C 2017 Unidirectional brain to muscle connectivity reveals motor cortex control of leg muscles during stereotyped walking *NeuroImage* **159** 403–16
[63] Liu J, Sheng Y and Liu H 2019 Corticomuscular coherence and its applications: a review *Front. Hum. Neurosci.* **13** 100
[64] Jungnickel E, Gehrke L, Klug M, Gramann K, Ayaz H and Dehais F 2019 MoBI—mobile brain/body imaging *Neuroergonomics* (New York: Academic) ch 10, pp 59–63
[65] Peterson S M and Ferris D P 2018 Differentiation in theta and beta electrocortical activity between visual and physical perturbations to walking and standing balance *eNeuro* **5** ENEURO.0207-18.2018
[66] He Y, Luu T P, Nathan K, Nakagome S and Contreras-Vidal J L 2018 A mobile brain-body imaging dataset recorded during treadmill walking with a brain-computer interface *Sci. Data* **5** 1–10
[67] Presacco A, Goodman R, Forrester L and Contreras-Vidal J L 2011 Neural decoding of treadmill walking from noninvasive electroencephalographic signals *J. Neurophysiol.* **106** 1875–87
[68] Wakeman D G and Henson R N 2015 A multi-subject, multi-modal human neuroimaging dataset *Sci. Data* **2** 1–10
[69] Hanada G 2018 Mobile brain and body imaging during walking motor tasks *PhD Thesis* University of Michigan
[70] Hollenstein N, Renggli C, Glaus B, Barrett M, Troendle M, Langer N and Zhang C 2021 Decoding EEG brain activity for multi-modal natural language processing (arXiv:2102.08655)
[71] Gravina R, Alinia P, Ghasemzadeh H and Fortino G 2017 Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges *Inf. Fusion* **35** 68–80
[72] Kartsch V J, Benatti S, Schiavone P D, Rossi D and Benini L 2018 A sensor fusion approach for drowsiness detection in wearable ultra-low-power systems *Inf. Fusion* **43** 66–76
[73] Ofner P, Schwarz A, Pereira J and Müller-Putz G R 2017 Upper limb movements can be decoded from the time-domain of low-frequency EEG *PLoS One* **12** 1–24
[74] Miller K J, Hermes D, Honey C J, Hebb A O, Ramsey N F, Knight R T, Ojemann J G and Fetz E E 2012 Human motor cortical activity is selectively phase-entrained on underlying rhythms *PLoS Comput. Biol.* **8** e1002655
[75] Asano Y M, Rupprecht C and Vedaldi A 2019 Self-labelling via simultaneous clustering and representation learning (arXiv:1911.05371)
[76] Cuturi M 2013 Sinkhorn distances: lightspeed computation of optimal transport *Advances in Neural Information Processing Systems* vol 26 pp 2292–300
[77] Singh S H, Peterson S M, Rao R P N and Brunton B W 2021 Mining naturalistic human behaviors in long-term video and neural recordings *J. Neurosci. Methods* **358** 109199
[78] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013
[79] Rosenberg A and Hirschberg J 2007 V-measure: a conditional entropy-based external cluster evaluation measure *Proc. 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* pp 410–20
[80] Han K, Vedaldi A and Zisserman A 2019 Learning to discover novel visual categories via deep transfer clustering *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 8401–9
[81] Friedman M 1937 The use of ranks to avoid the assumption of normality implicit in the analysis of variance *J. Am. Stat. Assoc.* **32** 675–701
[82] Conover W J 1998 *Practical Nonparametric Statistics* vol 350 (New York: Wiley)
[83] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300
[84] Tran Y, Thuraisingham R A, Wijesuriya N, Nguyen H T and Craig A 2007 Detecting neural changes during stress and fatigue effectively: a comparison of spectral analysis and sample entropy *2007 3rd Int. IEEE/EMBS Conf. on Neural Engineering* (IEEE) pp 350–3
[85] Baucom L B, Wedell D H, Wang J, Blitzer D N and Shinkareva S V 2012 Decoding the neural representation of affective states *NeuroImage* **59** 718–27
[86] Muraskin J, Brown T R, Walz J M, Tu T, Conroy B, Goldman R I and Sajda P 2018 A multimodal encoding model applied to imaging decision-related neural cascades in the human brain *NeuroImage* **180** 211–22
[87] Fatima S and Kamboh A M 2017 Decoding brain cognitive activity across subjects using multimodal M/EEG neuroimaging *2017 39th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 3224–7
[88] Gennaro F and de Bruin E D 2018 Assessing brain–muscle connectivity in human locomotion through mobile brain/body imaging: opportunities, pitfalls and future directions *Front. Public Health* **6** 39

[89] Peterson S M, Furuichi E and Ferris D P 2018 Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading *PLoS One* **13** e0200306
[90] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8
[91] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
[92] An K-N 1984 Kinematic analysis of human movement *Ann. Biomed. Eng.* **12** 585–97
[93] Fu W and Perry P O 2020 Estimating the number of clusters using cross-validation *J. Comput. Graph. Stat.* **29** 162–73
[94] Randazzo L, Iturrate I, Chavarriaga R, Leeb R and Millán J D R 2015 Detecting intention to grasp during reaching movements from EEG *2015 37th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 1115–8
[95] Alazrai R, Alwanni H and Daoud M I 2019 EEG-based BCI system for decoding finger movements within the same hand *Neurosci. Lett.* **698** 113–20
[96] Kringelbach M L and Deco G 2020 Brain states and transitions: insights from computational neuroscience *Cell Rep.* **32** 108128
[97] Pinto T P, Ramos M M R, Lemos T, Vargas C D and Imbiriba L A 2017 Is heart rate variability affected by distinct motor imagery strategies? *Physiol. Behav.* **177** 189–95
[98] Heremans E, Helsen W F and Feys P 2008 The eyes as a mirror of our thoughts: quantification of motor imagery of goal-directed movements through eye movement registration *Behav. Brain Res.* **187** 351–60
[99] Lebon F, Rouffet D, Collet C and Guillot A 2008 Modulation of EMG power spectrum frequency during motor imagery *Neurosci. Lett.* **435** 181–5
[100] Sani O G, Yang Y, Lee M B, Dawes H E, Chang E F and Shanechi M M 2018 Mood variations decoded from multi-site intracranial human brain activity *Nat. Biotechnol.* **36** 954–61
[101] Flesher S N, Downey J E, Weiss J M, Hughes C L, Herrera A J, Tyler-Kabara E C, Boninger M L, Collinger J L and Gaunt R A 2021 A brain-computer interface that evokes tactile sensations improves robotic arm control *Science* **372** 831–6
[102] Ravanelli M and Bengio Y 2018 Speaker recognition from raw waveform with SincNet *2018 IEEE Spoken Language Technology Workshop (SLT)* (IEEE) pp 1021–8
[103] Peterson S M, Singh S H, Dichter B, Scheid M, Rao R P N and Brunton B W 2022 AJILE12: long-term naturalistic human intracranial neural recordings and pose *Sci. Data* **9** 1–10
[104] Mani I and Zhang I 2003 kNN approach to unbalanced data distributions: a case study involving information extraction *Proc. Workshop on Learning From Imbalanced Datasets* vol 126 (ICML)
[105] Peterson Steven and Ferris Daniel 2021 Perturbed beam-walking task *OpenNeuro* (available at: https://openneuro.org/datasets/ds003739/) (https://doi.org/10.18112/openneuro.ds003739.v1.0.3)
[106] Peterson S M and Ferris D P 2021 Human electrocortical, electromyographical, ocular and kinematic data during perturbed walking and standing *Data Brief* **39** 107635