Extracting Human Face Similarity Judgments: Pairs or Triplets?

Linjie Li[1], Vicente Malave[2], Amanda Song[2], and Angela J. Yu[2]
(lil121@ucsd.edu, vmalave@cogsci.ucsd.edu, mas065@ad.ucsd.edu, ajyu@ucsd.edu)
[1]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA
[2]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA

Abstract

Two experimental protocols, pairwise rating and triplet ranking, have been commonly used for eliciting perceptual similarity judgments for faces and other objects. However, there has been little systematic comparison of the two methods. Pairwise rating has the advantage of greater precision, but triplet ranking is potentially a cognitive less taxing task, thus resulting in less noisy responses. Here, we introduce several information-theoretic measures of how useful responses from the two protocols are for the purpose of response prediction and parameter estimation. Using face similarity data collected on Amazon Mechanical Turk, we demonstrate that triplet ranking is significantly better for extracting subject-specific preferences, while the two are comparable when pooling across subjects. While the specific conclusions should be interpreted cautiously, due to the particularly simple Bayesian model for response generation utilized here, the work provides a information-theoretic framework for quantifying how repetitions within and across subjects can help to combat noise in human responses, as well as giving some insight into the nature of similarity representation and response noise in humans. More generally, this work demonstrates that substantial noise and inconsistency corrupt similarity judgments, both within- and across-subjects, with consequent implications for experimental design and data interpretation.
Keywords: similarity judgment, triplet ranking, pairwise rating, information theory, Bayesian modeling

Introduction

Several protocols have been developed in recent years to collect expensive and time-consuming human perceptual similarity judgments, such as among face images. Similarity is a pairwise numeric relationship between a pair of objects, where a higher value of similarity indicates that the objects are perceived to be more similar. For cognitive science, this is useful for predicting future judgments on unseen stimulus pairs, inferring a low-dimensional internal representation of the object space, identifying individual and group differences, and so on. For artificial intelligence, this type of data is often used as "ground truth" to label or categorize data, train or evaluate machine learning algorithms, predict future preferences in consumer marketing, etc.

There are two common ways to collect similarity ratings. Pairwise rating typically ask subjects to rate the perceived similarity of stimulus pairs using numbers on a specified numerical scale (such as a Likert scale). Algorithms such as classical multidimensional scaling (W. Torgerson, 1952; W. S. Torgerson, 1958) and modern variants (Borg & Groenen, 2005) make use of numeric, pairwise ratings. Another type of experiment has instead asked subjects to make ordinal judgments. One such algorithm, triplet ranking, consists of asking subjects to choose which pair of stimuli among three presented is the most similar. Relative comparisons were discussed early in the multidimensional scaling literature (W. S. Torgerson, 1958). Sometimes they are converted directly to numeric values and then used with an algorithm designed for pairs. More recently, algorithms have appeared in machine learning which learn directly from ordinal information Shepard (1962), or triplets, with no intermediate step: Generalized Non-metric Multidimensional Scaling (GNMDS) (Agarwal et al., 2007) , the Crowd Kernel algorithm (Tamuz et al., 2011) and Stochastic triplet Embedding (STE) (van der Maaten & Weinberger, 2012).

While both pairwise rating and triplet ranking have been used extensively in the literature, there has scant acknowledgement of the types of noise that can corrupt the two kinds of responses, and thus little systematic comparison of the informational utility of the two. In information terms, pairwise rating has the advantage of having more precision, and thus more capability of transmitting more information about human preferences. However, this greater precision could potentially be offset by the greater cognitive difficulty for humans to come up with numerical ratings, rather than making relative judgments. Moreover, different human subjects may interpret the numerical scale slightly differently, contributing to even more inter-subject noise. These factors can potentially translate into greater response noise or self-inconsistency, thus largely or even completely negating the precision advantage of pairwise rating over triplet ranking. A recent paper comparing several methods of collecting similarity data, (Demiralp et al. (2014)), compared pairwise rating in relative judgements terms of efficiency and consistency, and found that relative judgments can be more consistent.

More generally, in terms of the design of experiments involving extracting human similarity judgments, there has been little exploration of how many times a particular stimulus display (resulting in a judgment) should be repeated within subject or across subjects. Indeed, most algorithms simply ignore the fact there may be noise within- and across-subjects, treating the data is noise-free. In terms of experimental design, there is an obvious need to quantify and characterize the noise in order to chose the number of trials within and across subjects. In terms of cognitive science, a better understanding of the noise corrupting similarity judgments can yield insight into the nature of similarity representation in the brain.

In this paper, we utilize several different information-theoretic and probabilistic measures to quantify the information utility of pairwise rating and triplet ranking for extracting facial similarity judgments. Based on a simple Bayesian model, we compute the posterior distribution over the param-

eters of the distribution, as well as a marginal predictive distribution of the response for the next subject or the same subject on the next trial. We can then compute the information gained (entropy reduced) relative to both of these distributions, as well as a prediction error measure, for both both data collection methods. We collected face similarity judgment data, in both pairwise and triplet forms, on Amazon Mechanical Turk, and then used the various measures to quantify the informational utility of the two methods. We demonstrate that when predicting future responses from a subject's own data, triplet is better; however, when predicting a subject's response from others' data, the two methods are comparable.

The rest of the paper is organized as follows: in the Methods section, we explain the experimental design and the data modeling/analysis methods. In the Results section, we show how triplets are better than pairs within subject and vice versa across subjects. In the Discussion section, we discuss the broader impacts of our results, as well as fruitful directions of future work.

Methods

Experimental Design

We collected human similarity judgments on face images through Amazon Mechanical Turk, relying on two types of similarity judgments: a pairwise rating task, and a triplet ranking task. In the pairwise task, subjects were sequentially presented pairs of faces and asked to rate the similarity of each pair on a 9-point Likert scale (Figure 1). In the triplet task, we asked subjects to decide which pair of faces out of the three were the most similar to each other (Figure 2). The order within trials (which pairs were presented on left and right, which triplet appeared in which order), as well as the order of trials, was randomized for each subject. To ensure data quality, we also used catch trials, asking for ratings of identical stimuli (both in the pairwise and triplet cases) to identify non-compliant subjects.

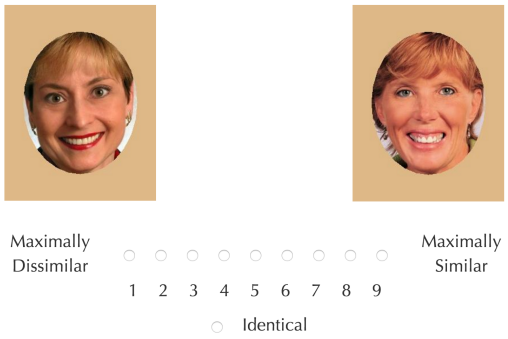

Figure 1: Sample questionnaire of pairwise rating.

In the experiment, we present 7 faces to 90 subjects in the two formats, exhaustively covering all possible pairs/triplets, for a total of 35 sets of triplets and 21 pairs. To aid the assessment of self-consistency, we present each pair and each triplet four times to each subject. Subjects carried out the

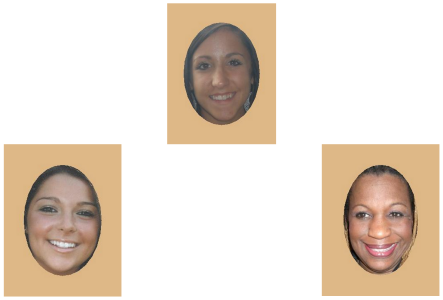

Figure 2: Sample questionnaire of triplet rankings.

experiments in five steps. Subjects are first presented a description of the task with an option of accepting it. Once the task was accepted, subjects complete a short training session, using an interface identical to the actual task interface. After the training session, subjects are prompted with the full set of faces and asked to think about the most similar and dissimilar stimuli in the set – this is to ensure subjects are aware of the full range of possible extent of similarity/dissimilarity, so as to reduce inconsistency on pairwise ratings. Afterward, subject complete the experimental task. In the last step, they provide information about themselves and submit their results. Two out of 90 subjects were thrown out due to being non-compliant.

The face images were taken from the 10k US Adult Faces database provided by Aude Oliva's group at MIT (Bainbridge et al., 2013) and then cropped for uniformity in presentation.

Data Conversion

Since subjects have 9 options in the pairwise setting but only 3 choices in the triple setting, direct comparison between these two types of data is difficult. Since the pairwise data can be easily converted to triplet data, while the reverse, or converting both to a common format, is not possible without making many assumptions about the underlying response generation process, we choose to convert the pairwise data into equivalent triplet data (based on which of the three pairs receives the highest similarity rating), and then use identical measures to compare the two in the remainder of the paper.

Correlation Analysis

As a first analysis of within-subject and across-subject consistency, we perform a correlation analysis. 35 stimuli (triplets) were rated by 88 subjects, each stimulus repeated 4 times, and there are 3 possible responses per trial.

We compute the average across-subject correlation as follows: for each stimulus, we compute the Pearson correlation coefficient between the empirical distributions (across the 3 possible responses) of two subjects, and average across all stimuli and all possible pairings of subjects.

We compute the average within-subject correlation as follows: for each stimulus in a subject, count the number of

pairwise responses that agree and then normalize by 6, the total number of pairwise comparisons across four trials. The score therefore has a value of 1/6 for really noisy data (e.g. (1, 1, 2)) and 1 for really consistent data (e. g. (3, 0, 0)). This value is then averaged across stimuli for each subject.

Statistical Modeling

Given a triplet composed of three faces {A,B,C}, the subject chooses which of the three pairs, {AB,AC,BC}, is most similar. For simplicity, we model a subject's responses to one stimulus as a multinomial distribution, $P(\mathbf{d}_x^i|vecr_x) = \text{Mult}(4,\mathbf{r}_x)$, where $d_x(l)$ is the number of times the l^{th} pair in a triplet x is chosen ($l = 1,2,3$), and $r_x l$ is the probability of choosing the l^{th} pair. In the across-subject analysis, we assume that all the subjects share the same preference vector $\mathbf{r}_x$ and thus generate responses from the same distribution.

We assume a conjugate prior, i.e. a Dirichlet prior distribution $p_0(\mathbf{r}_x^i;\theta)$, where $\theta = [1,1,1]$. The posterior distribution is thus also Dirichlet, and its mean is the predictive prior distribution for the next response/subject.

Information Gain Given the Bayesian response generation model, we can compare how much information is provided by pairwise or triplet data. We define *information gain* as the reduction in entropy, and we calculate the information gain relative to both the posterior distribution and the predictive distribution for the two methods, both within-subject and across subjects.

Prediction Error Use the Bayesian model, we can compute a predictive prior distribution over the next response/subject based on previous responses/subjects on the stimulus. We use MAP estimation (mode of the distribution) to predict the next response, and can therefore compute a predictive accuracy measure.

Results

As a first analysis of within-subject and across-subject consistency, we perform a correlation analysis in order to measure how consistent the subjects were with their own responses on previous trials, or with the responses of other subjects. Figure 3 shows that cross-correlation of responses across subjects for the same triplets, suggesting distinct subgroups of individuals based on overall similarity judgments. The clusters indicate that there are some consensus among the subjects that might be using the same criteria to judge facial similarity. As shown in Figure 4, we find that self-consistency is higher for triplet data than the pairwise data, while inter-subject consistency is higher for pairwise data than triplet data. The difference of across subject correlation between triplets and pairs is significant, while two methods have comparable within subject correlations.

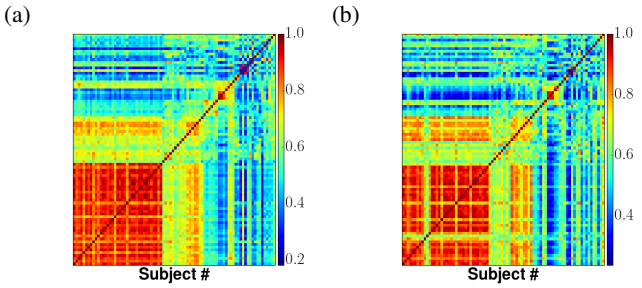

Figure 3: Cross correlation matrix of (a) triplet comparison and (b) pairwise rating. These heat maps indicate how subjects are correlated with each other. Subjects are ordered using hierarchical clustering.

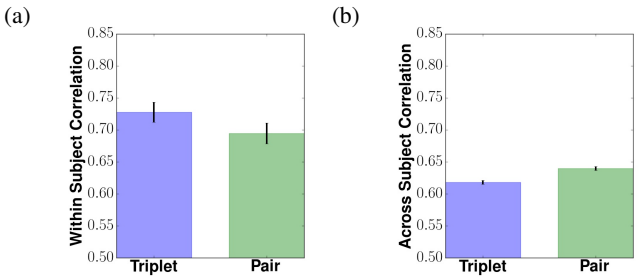

Figure 4: (a) Average within-subject correlation and (b) across-subject correlation for triplet comparison and pairwise rating. Error bars denote standard errors of the mean.

Correlation (consistency) is a very coarse measure of the utility of data, as a subject giving the same response (e.g. 1) to all stimuli on all trials would achieve maximal correlation but actually yield minimal information about any true preferences. We therefore need measures that quantify not only consistency but also diversity of responses, and that brings us naturally to information-theory. We therefore utilize a simple Bayesian generative model to capture how noisy responses arise from true (hidden) similarity percepts. We then use this model to compute both entropy-reduction related to the model parameters, based on the posterior distribution, as well as to the subject's future responses, based on the predictive prior distribution (see Methods).

We find that triplets are more informative than pairs within subject, while they are comparable across subjects. Figure 5 presents the information gain related to the posterior distribution and the predictive distribution, as we see more and more data from one subject. After all four repetitions, the information provided by triplet comparison is more than that gained from pairwise ranking, and the difference is significant. Even though the incremental information gain is decreasing as the number of repetitions increases (as we would expect), there is still significant information gained through the fourth repetition.

The result of the across-subjects analysis (Figure 6) indicates that the two methods are similar in efficacy. The incremental information gain converges to 0 as number of subjects increases. However, the first 15 subjects provide around 80% of total predictive information gain and approximately 60% of total posterior information gain.

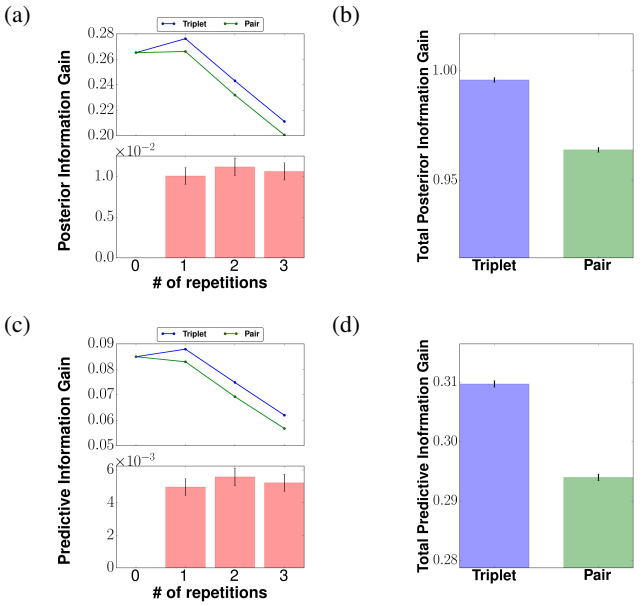

Figure 5: Within-subject information gain of (a) posterior distribution and (c) predictive distribution; total within-subject information gain of (b) posterior distribution and (d) predictive distribution. The lower bar graphs in (a) and (c) plot the point-wise differences of the upper line plots (triplet - pairwise).

While information gain measures how confident we are about the estimated model parameters or about the model predictions of future responses, it does not tell us how much better the model is getting at predicting future responses. In particular, while model precision can improve, its accuracy may saturate or even decline. One can think of the entropy-related measures as quantifying *variance* in predictive accuracy, while the prediction error measure quantifes *bias*. We therefore also use the Dirichlet-multinomial response model to make predictions (using MAP estimate) and compute an average accuracy measure by comparing to human responses.

As illustrated in Figure 7, as more and more data are collected, both within- and across-subject prediction error and entropy decrease. Prediction error decreases sharply before the 5^{th} subjects across subjects and the second repetition within subjects. Notably, predictive error rate stops decreasing sooner than predictive entropy, indicating the model is probably somewhat misspecified: as predictive uncertainty (variance) decreases, total accuracy is already saturated (bias persists). In addition, prediction error converges to approximately 0.23 within subject, compared to 0.32 across subjects.

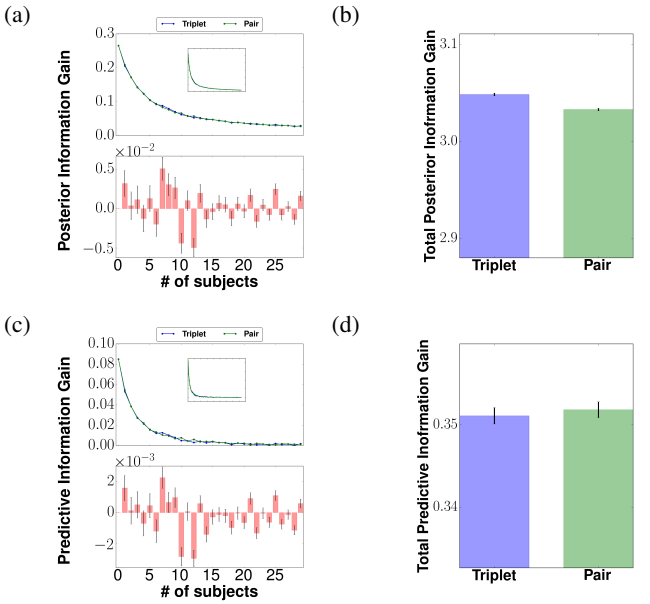

Figure 6: Across-subjects information gain of (a) posterior distribution and (c) predictive distribution; cumulative across subjects information gain of (b) posterior distribution and (d) predictive distribution. The upper line plots in (a) and (c) show the results from the first 30 subjects; the insets show the results from all subjects. The lower bar graphs in (a) and (c) show the point-wise differences of the upper line plots (triplet - pairwise).

This indicates that variability across subjects is not only due to noise but also due to systematic individual differences in preferences.

Similar to other measures, triplet data yield a higher predictive accuracy within-subject than pairwise data, but are similar when predicting across subjects.

Discussion

Similarity learning has been a well-studied topic in cognitive science research. In this area, the study of facial similarity has been particularly prominent, both due to the important role facial processing plays in human interactions, as well as due to the extremely high dimensionality of the face image space and lack of an obvious low-dimensional featural representation. While various experimental methodologies have been utilized to elicit facial similarity judgments, there has been little systematic comparison of their efficacy. More troublingly, most algorithms have assumed human responses to be free of noise and to be completely interchangeable from one subject's response to another's. To tackle some of these issues, we introduce a suite of statistical and information theoretic measures for investigating the amount and type of noise within- and across-subjects. We applied these methods, along with a simple Dirichlet-multinomial Bayesian model for response generation, to a novel crowdsourced dataset. We found that triplet ranking is more informative and predictive

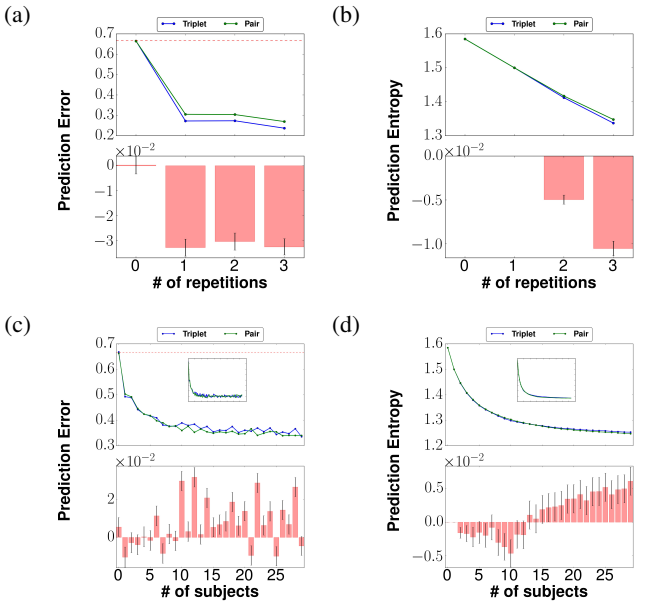

Figure 7: Within-subject (a) prediction error and (b) prediction entropy; across-subject (c) prediction error and (d) prediction entropy. Lower bar graphs indicates the point-wise differences of the upper line plots (triplet - pairwise).

for eliciting facial similarity judgments within a particular subject. It is consistent with the hypothesis that humans are better at making relative similarity judgments. Although pairwise rating has greater granularity, it has often been observed that humans give more self-consistent responses when reporting relative preferences than assigning numeric values to individual items, especially in complex judgments that involve high-dimensional input. Forcing humans to assign numerical values to complex judgments can not only fail to add information but can additionally corrupt the information available in simpler relative ranking responses. In contrast to within-subject analysis, we found triplet ranking and pairwise rating to be comparable in across-subject modeling. Why this is the case is an interesting topic for future investigations.

We covered two of the possible experimental designs used to collect similarity data, but other exist, such as spatial arrangement (Demiralp et al. (2014)), which could also benefit from our analysis. Our model has its limitations as well. We convert pairwise ratings into equivalent triplet rankings for all our analysis: this step may induce a loss of information in the pairwise data. Converting pairs to triplets not only allows an apple to apple comparison, but also minimizes the number of assumptions we need to make. One may argue that measuring mutual information between subjects' responses is a more intuitive and reasonable model free comparison. However, with 9 possible choices in the pairwise setting, we need much more data to compute mutual information of the pairwise data than the triplet data, which make it infeasible.

Our results have broader implications. Our analysis is relevant to the general problem of crowdsourcing similarity models. Depending on the goal of a similarity experiment, different methods should be chosen. If the experiment aims at modeling personal preferences, triplet comparison appear to provide higher quality data. When the goal is to find a population-level model of similarity judgements, without worrying about individual differences, then pairwise data compare well to triplet data. More generally, our work speaks to how many repetitions within and across subjects should be employed. According to the information gain analysis, our data reveal that most of the information is provided by the first 15 subjects. However, as any given model is probably not perfect in capturing human responses, greater model precision may not translate to greater ability to predict future human responses. This is consistent with our observation that, in terms of predicting future responses, two repetitions within subject and five repetitions across subjects exhaust information gain. Moreover, we provided a framework to compute how good a data collection scheme is.

The simple Dirichlet-multinomial model we introduced provides a baseline for the comparison between triplet rankings and pairwise ratings. If a more complicated model is to be proposed in the future, with our framework, all the analysis can be performed to quantify the efficacy of a technique, or compare across various techniques. An obvious next step for a better model is to integrate the relationship among different stimuli. The response model we utilized here simply assumes all triplets to be independent from each other. Similarity data are usually used to fit a latent variable model (such as multidimensional scaling), where faces are shared among triplets, and therefore one can use one set of triplet responses to potentially predict responses about another triplet with partially overlapping faces, or even completely new data. One valuable direction of future research would be to find a low-dimensional embedding of face images, in which we model similarity responses as arising from the perceptual distance between faces. The analyses proposed in this paper easily extend to latent variable models, and would be the focus of our future work.

Acknowledgments

This research is in part funded by a UCSD FISP awarded to AJY. VM is supported by a San Diego Fellowship.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D. J., & Belongie, S. (2007). Generalized non-metric multidimensional scaling. In *International conference on artificial intelligence and statistics* (pp. 11–18).

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Cao, C., & Ai, H.-Z. (2015). Facial similarity learning with humans in the loop. *Journal of Computer Science and Technology*, *30*(3), 499–510.

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Navarro, D. J., & Griffiths, T. L. (2006). A nonparametric bayesian method for inferring features from similarity judgments. In *Advances in neural information processing systems* (pp. 1033–1040).

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Torgerson, W. (1952). Multidimensional Scaling: I. Theory and Method*. *Psychometrika*, *17*(4), 401–419.

Torgerson, W. S. (1958). Theory and methods of scaling.

van der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding.