MAP IT TO VISUALIZE REPRESENTATIONS SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

1 FURTHER PERSPECTIVES ON t-SNE, UMAP, PACMAP, AND VARIANTS

As mentioned in Section 1 (Introduction) of the main paper, a large plethora of dimensionality reduction methods exists and an excellent repository for more information is e.g. `https://jlmelville.github.io/smallvis/`. In this paper, following recent literature, the main algorithms are considered to be t-SNE (van der Maaten, 2014), and UMAP (McInnes et al., 2020), and we include the empirically motivated PacMap (Wang et al., 2021). For context, TriMap (Amid & Warmuth, 2019) and LargeVis (Tang et al., 2016) are discussed below. The t-SNE theory has been outlined in Section 2 of the main paper.

In the paper introducing UMAP, McInnes et al. (2020) argues that t-SNE should be considered the current state-of-the-art at that time, and mentions computational scalability as a main benefit of UMAP versus t-SNE. The main aspect with respect to computational scalability for UMAP versus t-SNE is that the simplical set theory shows that for UMAP normalization over pairwise similarities (probabilities in t-SNE) is not needed, as opposed to t-SNE. This illustrates the importance of the sound theoretical foundation of UMAP. As further described in (McInnes et al., 2020), UMAP's simplical set cross-entropy cost function resembles in several ways the LargeVis (Tang et al., 2016) cost function. LargeVis also avoids normalization in the embedding space, albeit from a more heuristic point of view, but not in the input space where a procedure similar to the one used in Barnes-Hut t-SNE (van der Maaten, 2014) is used. Avoiding normalization in the embedding space is key to the negative sampling strategy employed in LargeVis and which is key to its computational scalability, also an integral component in the UMAP optimization. LargeVis is not included in the experimental part of this paper to avoid clutter, but is an influential algorithm in the t-SNE family. McInnes et al. (2020) and Wang et al. (2021), for instance, have both extensive comparative experiments sections also involving LargeVis.

A motivation for Wang et al. (2021) is to discuss preservation of local structure versus global structure. They propose a heuristic method, PacMap, which is intended to strike a balance between TriMap (Amid & Warmuth, 2019) (better at preserving global structure) and t-SNE/UMAP (local structure). TriMap is is a triplet loss-based method. Wang et al. (2021) argues that TriMap is the first successful triplet constraint method (as opposed to (Hadsell et al., 2006; van der Maaten & Weinberger, 2012; Wilber et al., 2015)) but claims that without PCA initialization ”TriMap's global structure is ruined”. PacMap is based on a study of the principles behind attractive and repulsive forces and finds that forces should be exerted on further points and sets up a heuristically designed procedure for treating near pairs, mid-near pairs, and non-neighbors.

Understanding t-SNE versus UMAP, in particular, from a theoretical perspective, has gained interest in the recent years. Damrich & Hamprecht (2021) studies the interplay between attractive and repulsive forces in UMAP in detail and comes to the conclusion that UMAP is actually not exactly optimizing the cost function put forth in (McInnes et al., 2020). Bohm et al. (2020) studies the whole attraction-repulsion spectrum and find cases where UMAP may diverge.

Kobak & Linderman (2021) show that UMAP's initialization (Laplacian eigenmaps (Belkin & Niyogi, 2003)) is very important for UMAP's results and claims that t-SNE can be improved by a similar initialization. Wang et al. (2021) also studies initialization, and claim that both UMAP but also TriMap are very dependent on initialization. A further comment on a relationship between t-SNE and Laplacian eigenmaps is provided in Section 2 of this Supplementary material. Draganov et al. (2023) argues that the normalization aspect is basically the key difference between t-SNE and UMAP, and suggests a way to toggle between the two approaches.

In this paper, sampling is also demonstrated (Figure 8 in the main paper, Section 3 in this Supplementary), sharing some similarity to the sampling strategies invoked in e.g. LargeVis and UMAP.

It should also be mentioned that dimensionality reduction methods inspired by the t-SNE approach by alternative divergences to the Kullback-Leibler over joint pairwise probabilities have been studied to some degree (Bunte et al., 2012; Naryan et al., 2015; Huang et al., 2022). However, these works have not discussed projective properties of divergence measures and have not contributed to understanding the aspect of normalization.

2 Propositions with comments

For the benefit of the reader, the well-known Kullback-Leibler-based t-SNE relation from the main paper, which van der Maaten & Hinton (2008) builds on, is proved:

$$\operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} KL(P||Q) = \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} \sum_{i,j} p_{ij}\log\frac{p_{ij}}{q_{ij}} = \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} \sum_{i,j} p_{ij}\log q_{ij}. \tag{1}$$

Proof.

$$\operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} KL(P||Q) = \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} \sum_{i,j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{2}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} \underbrace{\sum_{i,j} p_{ij}\log p_{ij}}_{\text{constant}} - \sum_{i,j} p_{ij}\log q_{ij}. \tag{3}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\sum_{i,j} p_{ij}\log q_{ij}. \tag{4}$$

□

Proposition 1. *[Minimizing $KL(P||Q)$ and the role of normalization]. Let $q_{ij} = \frac{\tilde{q}_{ij}}{\sum_{n,m}\tilde{q}_{nm}}$. Then*

$$\operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} KL(P||Q) = \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\sum_{i,j} p_{ij}\log\tilde{q}_{ij} + \log\sum_{n,m}\tilde{q}_{nm}. \tag{5}$$

Proof.

$$\operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} KL(P||Q) = \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\sum_{i,j} p_{ij}\log q_{ij} \tag{6}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\sum_{i,j} p_{ij}\log\frac{\tilde{q}_{ij}}{\sum\limits_{n,m}\tilde{q}_{nm}} \tag{7}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\left(\sum_{i,j} p_{ij}\log\tilde{q}_{ij} - \underbrace{\sum_{i,j} p_{ij}}_{=1}\log\sum_{n,m}\tilde{q}_{nm}\right) \tag{8}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d} -\sum_{i,j} p_{ij}\log\tilde{q}_{ij} + \log\sum_{n,m}\tilde{q}_{nm}. \tag{9}$$

□

Comment to Proposition 1. Deriving the t-SNE cost function without first expressing q_{ij}[1] in a particular form (t-distribution (Cauchy distribution) or Gaussian) shows a general property of the cost function which has received little or no attention in the previous literature. The t-SNE cost function is by the above result expressed by two terms where the first term relates the normalized p_{ij} to unnormalized $\tilde{q}_{ij}$. The second term only involves the unnormalized $\tilde{q}_{ij}$.

From this expression, an interesting aspect of the t-SNE cost function, which seems not to have been discussed much in the literature, is an intrinsic connection to Laplacian eigenmaps (Belkin & Niyogi, 2003).

In Laplacian eigenmaps, the aim is to find a low-dimensional embedding $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}_d$ from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}_D$. It is assumed that some similarity measure can be defined in the input space, pairwise over $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which can be denoted w_{ij}. The Laplacian cost function is essentially $\sum_{i,j} w_{ij} ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2$ to be minimized over $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}_d$, however given orthogonality constraints on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}_d$ to avoid trivial minima. Since for t-SNE, $\tilde{q}_{ij}$ is a function of $||\boldsymbol{z}_i - \boldsymbol{z}_j||^2$ for both the t-distribution (Cauchy distribution) and the Gaussian, there is a close link between t-SNE and Laplacian eigenmaps. For instance, for $\tilde{q}_{ij} = [1 + ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2]^{-1}$, the first term becomes $\sum_{i,j} w_{ij} \log ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2$ and for the Gaussian the corresponding term becomes $\sum_{i,j} w_{ij} ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2$ up to a proportionality constant (this was also pointed out in (Trosten et al., 2023)). By letting $w_{ij} = p_{ij}$ the link becomes obvious. In t-SNE, there are no orthogonality constraints on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}_d$. Instead, trivial solutions are avoided by the constraint posed by $\sum_{n,m} \tilde{q}_{nm}$.

Recently, Cai & Ma (2022) provided an elaborate spectral analysis concluding that t-SNE is connected to Laplacian eigenmaps. By the above analysis, this is evident from the form of the t-SNE cost function itself.

Proposition 2. *[Gradient of $KL(P||Q)$]*

$$\frac{\partial}{\partial \boldsymbol{z}_i} KL(P||Q) = -4 \sum_j \left(p_{ij} - q_{ij}\right) \tilde{q}_{ij} (\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{10}$$

Proof.

$$\frac{\partial}{\partial \boldsymbol{z}_i} KL(P||Q) = \frac{\partial}{\partial \boldsymbol{z}_i} \underbrace{-\sum_{i,j} p_{ij} \log \tilde{q}_{ij} + \log \sum_{n,m} \tilde{q}_{nm}}_{\overset{\text{def}}{=} C}. \tag{11}$$

The following derivation resembles (van der Maaten & Hinton, 2008). Note that if $\boldsymbol{z}_i$ changes, the only pairwise distances that change are d_{ij} and d_{ji} where $d_{ij} = ||\boldsymbol{z}_i - \boldsymbol{z}_j||$. Hence, the gradient of the cost function C with respect to $\boldsymbol{z}_i$ is given by

$$\frac{\partial}{\partial \boldsymbol{z}_i} C = \sum_j \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}} \right) (\boldsymbol{z}_i - \boldsymbol{z}_j) = 2 \sum_j \frac{\partial C}{\partial d_{ij}} (\boldsymbol{z}_i - \boldsymbol{z}_j). \tag{12}$$

Furthermore

$$\frac{\partial C}{\partial d_{ij}} = -\sum_{k,l} p_{kl} \frac{\partial}{\partial d_{ij}} \log \tilde{q}_{kl} + \frac{\partial}{\partial d_{ij}} \log \sum_{n,m} \tilde{q}_{nm} \tag{13}$$

$$= -\sum_{k,l} p_{kl} \frac{1}{\tilde{q}_{kl}} \frac{\partial}{\partial d_{ij}} \tilde{q}_{kl} + \frac{1}{\sum\limits_{n,m} \tilde{q}_{nm}} \sum_{n,m} \frac{\partial}{\partial d_{ij}} \tilde{q}_{nm}. \tag{14}$$

The gradient is only non-zero for $k = i$, $l = j$ and for $n = i$, $m = j$, yielding

$$\frac{\partial C}{\partial d_{ij}} = -p_{ij} \frac{1}{\tilde{q}_{ij}} \frac{\partial}{\partial d_{ij}} \tilde{q}_{ij} + \frac{1}{\sum\limits_{n,m} \tilde{q}_{nm}} \frac{\partial}{\partial d_{ij}} \tilde{q}_{ij}. \tag{15}$$

[1] In the original SNE paper (Hinton & Roweis, 2002) the Gaussian distribution was used $q_{ij} = \exp(-\kappa ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2) / \sum_{n,m} \exp(-\kappa ||\boldsymbol{z}_n - \boldsymbol{z}_m||^2)$. The argument in (van der Maaten & Hinton, 2008) was that the t-distribution helps mitigate the so-called crowding problem. It is possible to formulate the joint probabilities as functions of other distance functions than the Euclidean or in terms of similarity measures.

Note that for $\tilde{q}_{ij} = \left[\frac{1}{1+||\boldsymbol{z}_i-\boldsymbol{z}_j||^2}\right]$, which is the t-distribution, or alternatively for $\tilde{q}_{ij} = \exp(-\kappa||\boldsymbol{z}_i-\boldsymbol{z}_j||^2)$, the Gaussian distribution, we have

$$\frac{\partial}{\partial d_{ij}}\tilde{q}_{ij} = 2\tilde{q}_{ij}^2. \tag{16}$$

Hence,

$$\frac{\partial C}{\partial d_{ij}} = 2\left(-p_{ij}\tilde{q}_{ij} + \frac{1}{\sum\limits_{n,m}\tilde{q}_{nm}}\tilde{q}_{ij}^2\right) = 2\left(-p_{ij}+q_{ij}\right)\tilde{q}_{ij}, \tag{17}$$

since $q_{ij} = \frac{\tilde{q}_{ij}}{\sum\limits_{n,m}\tilde{q}_{nm}}$. Finally, this yields

$$\frac{\partial}{\partial \boldsymbol{z}_i}C = \frac{\partial}{\partial \boldsymbol{z}_i}KL(P||Q) = -4\sum_j\left(p_{ij}-q_{ij}\right)\tilde{q}_{ij}(\boldsymbol{z}_j-\boldsymbol{z}_i). \tag{18}$$

□

Proposition 3. *[Minimizing $CS(P_m||Q_m)$ with respect to $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d$]*

$$\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; CS(\tilde{P}_m||\tilde{Q}_m) = \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; -\log\sum_j\tilde{p}_j\tilde{q}_j + \frac{1}{2}\log\sum_j\tilde{q}_j^2. \tag{19}$$

Proof.

$$\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; CS(P_m||Q_m) \tag{20}$$

$$= \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; -\log\frac{\sum\limits_j\tilde{p}_j\tilde{q}_j}{\underbrace{\left(\sum\limits_j\tilde{p}_j^2\right)^{\frac{1}{2}}}_{\text{independent of }\boldsymbol{z}}\left(\sum\limits_j\tilde{q}_j^2\right)^{\frac{1}{2}}} \tag{21}$$

$$= \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; -\log\sum_j\tilde{p}_j\tilde{q}_j + \frac{1}{2}\log\sum_j\tilde{q}_j^2. \tag{22}$$

$$\tag{23}$$

□

Proposition 4. *[Gradient of $CS(P_m||Q_m)$]*

$$\frac{\partial}{\partial \boldsymbol{z}_i}CS(P_m||Q_m) = -4\sum_j\left[\frac{\tilde{p}_j}{\sum\limits_{j'}\tilde{p}_{j'}\tilde{q}_{j'}} - \frac{\tilde{q}_j}{\sum\limits_{j'}\tilde{q}_{j'}^2}\right]\tilde{q}_{ij}^2(\boldsymbol{z}_j-\boldsymbol{z}_i). \tag{24}$$

Proof.

$$\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; CS(P_m||Q_m) = \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\in\mathbb{R}_d}{\arg\min}\; -\log\sum_j\tilde{p}_j\tilde{q}_j + \frac{1}{2}\log\sum_j\tilde{q}_j^2 \tag{25}$$

$$\tag{26}$$

There are several ways to proceed. Here, it is chosen to start by expressing $CS(P_m||Q_m)$ explicitly into cross-product terms $\tilde{p}_{jk'}\tilde{q}_{jk}$. For convenience, the derivation is split into two parts.

$$\frac{\partial}{\partial \boldsymbol{z}_i} - \log\sum_j\tilde{p}_j\tilde{q}_j = -\frac{1}{\sum\limits_{j'}\tilde{p}_{j'}\tilde{q}_{j'}}\frac{\partial}{\partial \boldsymbol{z}_i}\sum_j\tilde{p}_j\tilde{q}_j = -\frac{1}{\sum\limits_{j'}\tilde{p}_{j'}\tilde{q}_{j'}}\frac{\partial}{\partial \boldsymbol{z}_i}\sum_j\sum_{k',k}\tilde{p}_{jk'}\tilde{q}_{jk}. \tag{27}$$

Look first at the case $j \neq i$. Then $\frac{\partial}{\partial \boldsymbol{z}_i} \sum_{k',k} \tilde{p}_{jk'}\tilde{q}_{jk}$ will have non-zero terms for $k = i$, hence

$$\frac{\partial}{\partial \boldsymbol{z}_i} \sum_{k',k} \tilde{p}_{jk'}\tilde{q}_{jk} = \sum_{k'} \tilde{p}_{jk'} \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ji} = \tilde{p}_j \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ji}. \tag{28}$$

For $j = i$,

$$\frac{\partial}{\partial \boldsymbol{z}_i} \sum_{k',k} \tilde{p}_{ik'}\tilde{q}_{ik} = \sum_{k} \left(\sum_{k'} \tilde{p}_{ik'} \right) \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ik} = \sum_{k} \tilde{p}_k \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ik}. \tag{29}$$

Hence,

$$\frac{\partial}{\partial \boldsymbol{z}_i} \sum_{j} \sum_{k',k} \tilde{p}_{jk'}\tilde{q}_{jk} = \sum_{j,j\neq i} 2\tilde{p}_j \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ij}. \tag{30}$$

Note that for $\tilde{q}_{ij} = \left[\frac{1}{1+||\boldsymbol{z}_i - \boldsymbol{z}_j||^2}\right]$, which is the t-distribution, or alternatively for $\tilde{q}_{ij} = \exp(-\kappa||\boldsymbol{z}_i - \boldsymbol{z}_j||^2)$, the Gaussian distribution, we have

$$\frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_{ij} = 2\tilde{q}_{ij}^2(\boldsymbol{z}_j - \boldsymbol{z}_i) \tag{31}$$

Thus

$$\frac{\partial}{\partial \boldsymbol{z}_i} - \log \sum_{j} \tilde{p}_j \tilde{q}_j = -4 \frac{1}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} \sum_{j,j\neq i} \tilde{p}_j \tilde{q}_{ij}^2 (\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{32}$$

Alternatively,

$$\frac{\partial}{\partial \boldsymbol{z}_i} - \log \sum_{j} \tilde{p}_j \tilde{q}_j = - \frac{1}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} \frac{\partial}{\partial \boldsymbol{z}_i} \sum_{i} \tilde{p}_i \tilde{q}_i = - \frac{1}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} \sum_{j} \tilde{p}_j \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_j \tag{33}$$

and then work with $\tilde{q}_j = \sum\limits_{k'} \tilde{q}_{jk'}$. For the second part, consider

$$\frac{\partial}{\partial \boldsymbol{z}_i} \frac{1}{2} \log \sum_{j} \tilde{q}_j^2 = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}_{j'}^2} \frac{\partial}{\partial \boldsymbol{z}_i} \sum_{i} \tilde{q}_i^2 = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}_{j'}^2} \frac{\partial}{\partial \boldsymbol{z}_i} \sum_{j} \sum_{k',k} \tilde{q}_{jk'}\tilde{q}_{jk}. \tag{34}$$

and work in a similar fashion as above from there, or express

$$\frac{\partial}{\partial \boldsymbol{z}_i} \frac{1}{2} \log \sum_{j} \tilde{q}_j^2 = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}_{j'}^2} \frac{\partial}{\partial \boldsymbol{z}_i} \sum_{i} \tilde{q}_i^2 = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}_{j'}^2} \sum_{j} 2\tilde{q}_j \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}_j \tag{35}$$

and insert $\tilde{q}_j = \sum\limits_{k'} \tilde{q}_{jk'}$. This gives

$$\frac{\partial}{\partial \boldsymbol{z}_i} \frac{1}{2} \log \sum_{j} \tilde{q}_j^2 = 4 \frac{1}{\sum\limits_{j'} \tilde{q}_{j'}^2} \sum_{j,j\neq i} \tilde{q}_j \tilde{q}_{ij}^2 (\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{36}$$

Hence, when taken together:

$$\frac{\partial}{\partial \boldsymbol{z}_i} CS(P_m||Q_m) = -4 \sum_{j,j\neq i} \left[\frac{\tilde{p}_j}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} - \frac{\tilde{q}_j}{\sum\limits_{j'} \tilde{q}_{j'}^2} \right] \tilde{q}_{ij}^2 (\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{37}$$

□

Proposition 5. *[The* map *CS divergence is projective]. Let $\boldsymbol{z} = m(\boldsymbol{x}) + \varepsilon$ and assume $p(\boldsymbol{x}) = \frac{\tilde{p}(\boldsymbol{x})}{Z_p}$ and $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$ where $\tilde{p}(\boldsymbol{x})$ and $\tilde{q}(\boldsymbol{z})$ are unnormalized with Z_p and Z_q as the respective normalization constants. Then*

$$CS(p(\boldsymbol{x})||q(\boldsymbol{z})) = CS(\tilde{p}(\boldsymbol{x})||\tilde{q}(\boldsymbol{z})). \tag{38}$$

Proof.

$$\begin{aligned}
&CS(p(\boldsymbol{x})||q(\boldsymbol{z})) && (39)\\
&= -\log \frac{\int\int \frac{\tilde{p}(\boldsymbol{x})}{Z_p}\frac{\tilde{q}(\boldsymbol{z})}{Z_q} f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}}{\left(\int\int \left(\frac{\tilde{p}(\boldsymbol{x})}{Z_p}\right)^2 f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}} \left(\int\int \left(\frac{\tilde{q}(\boldsymbol{z})}{Z_q}\right)^2 f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}}} && (40)\\
&= -\log \frac{\frac{1}{Z_p Z_q}\int\int \tilde{p}(\boldsymbol{x})\tilde{q}(\boldsymbol{z}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}}{\left(\frac{1}{Z_p^2}\int\int \tilde{p}^2(\boldsymbol{x}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}} \left(\frac{1}{Z_q^2}\int\int \tilde{q}^2(\boldsymbol{z}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}}} && (41)\\
&= -\log \frac{\int\int \tilde{p}(\boldsymbol{x})\tilde{q}(\boldsymbol{z}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}}{\left(\int\int \tilde{p}^2(\boldsymbol{x}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}} \left(\int\int \tilde{q}^2(\boldsymbol{z}) f(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}\right)^{\frac{1}{2}}} && (42)\\
&= CS(\tilde{p}(\boldsymbol{x})||\tilde{q}(\boldsymbol{z})). && (43)
\end{aligned}$$

□

Proposition 6. *[Minimizing $\widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z}))$ with respect to $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}_d$]*

$$\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ \widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ -\log \sum_j \tilde{p}(\boldsymbol{x}_j)\tilde{q}(\boldsymbol{z}_j) + \frac{1}{2}\log \sum_j \tilde{q}^2(\boldsymbol{z}_j). \tag{44}$$

Proof.

$$\begin{aligned}
&\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ CS(p(\boldsymbol{x})||q(\boldsymbol{z})) && (45)\\
&= \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ -\log \frac{\sum_j \tilde{p}(\boldsymbol{x}_j)\tilde{q}(\boldsymbol{z}_j)}{\underbrace{\left(\sum_j \tilde{p}^2(\boldsymbol{x}_j)\right)^{\frac{1}{2}}}_{\text{independent of } \boldsymbol{z}} \left(\sum_j \tilde{q}^2(\boldsymbol{z}_j)\right)^{\frac{1}{2}}} && (46)\\
&= \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ -\log \sum_j \tilde{p}(\boldsymbol{x})\tilde{q}(\boldsymbol{z}_j) + \frac{1}{2}\log \sum_j \tilde{q}^2(\boldsymbol{z}_j). && (47)\\
& && (48)
\end{aligned}$$

□

Proposition 7. *[Gradient of $\widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z}))$]*

$$\frac{\partial}{\partial \boldsymbol{z}_i}\widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = -\sum_j \left[\frac{\tilde{p}(\boldsymbol{x}_j)}{\sum_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} - \frac{\tilde{q}(\boldsymbol{z}_j)}{\sum_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})}\right] \frac{\partial}{\partial \boldsymbol{z}_i}\tilde{q}(\boldsymbol{z}_j). \tag{49}$$

Proof.

$$\underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ CS(p(\boldsymbol{x})||q(\boldsymbol{z})) = \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \in \mathbb{R}_d}{\arg\min}\ -\log \sum_j \tilde{p}(\boldsymbol{x}_j)\tilde{q}(\boldsymbol{z}_j) + \frac{1}{2}\log \sum_j \tilde{q}^2(\boldsymbol{z}_j). \tag{50}$$

The derivation is split into two parts. First,

$$\frac{\partial}{\partial \boldsymbol{z}_i} - \log \sum_j \tilde{p}(\boldsymbol{x}_j)\tilde{q}(\boldsymbol{z}_j) = -\frac{1}{\sum\limits_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} \sum_j \tilde{p}(\boldsymbol{x}_j) \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}(\boldsymbol{z}_j). \tag{51}$$

Second,

$$\frac{\partial}{\partial \boldsymbol{z}_i} \frac{1}{2} \log \sum_j \tilde{q}^2(\boldsymbol{z}_j) = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \sum_j \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}^2(\boldsymbol{z}_j) = \frac{1}{2} \frac{1}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \sum_j 2\tilde{q}(\boldsymbol{z}_j) \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}(\boldsymbol{z}_j). \tag{52}$$

Taken together, thus

$$\frac{\partial}{\partial \boldsymbol{z}_i} \widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = -\sum_j \left[\frac{\tilde{p}(\boldsymbol{x}_j)}{\sum\limits_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} - \frac{\tilde{q}(\boldsymbol{z}_j)}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \right] \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}(\boldsymbol{z}_j). \tag{53}$$

□

Proposition 8. *[Gradient of $\widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z}))$ with kernel smoothing]. Let $\hat{\tilde{q}}(\boldsymbol{z}_j) = \sum_k \kappa_z(\boldsymbol{z}_j - \boldsymbol{z}_k)$ and $\hat{\tilde{p}}(\boldsymbol{x}_j) = \sum_k \kappa_p(\boldsymbol{x}_j - \boldsymbol{x}_k)$ for shift-invariant kernel functions $\kappa_z(\cdot)$ and $\kappa_p(\cdot)$. Then*

$$\frac{\partial}{\partial \boldsymbol{z}_i} \widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = -4 \sum_j \left[\frac{\tilde{p}_j}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} - \frac{\tilde{q}_j}{\sum\limits_{j'} \tilde{q}^2_{j'}} \right] \kappa_z^2(\boldsymbol{z}_j - \boldsymbol{z}_i)(\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{54}$$

Proof. A shift-invariant kernel function satisfies $\kappa(\boldsymbol{z}_j - \boldsymbol{z}_i) = \kappa(d_{ij})$ where $d_{ij} = ||\boldsymbol{z}_j - \boldsymbol{z}_i||^2$. Hence, similar to the derivation for Proposition 2, we will have

$$\frac{\partial}{\partial \boldsymbol{z}_i} \widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = -\sum_j \left[\frac{\tilde{p}(\boldsymbol{x}_j)}{\sum\limits_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} - \frac{\tilde{q}(\boldsymbol{z}_j)}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \right] \frac{\partial}{\partial \boldsymbol{z}_i} \tilde{q}(\boldsymbol{z}_j) \tag{55}$$

$$= -\sum_j \left[\frac{\tilde{p}(\boldsymbol{x}_j)}{\sum\limits_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} - \frac{\tilde{q}(\boldsymbol{z}_j)}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \right] \left(\frac{\partial \tilde{q}(\boldsymbol{z}_j)}{\partial d_{ij}} + \frac{\partial \tilde{q}(\boldsymbol{z}_j)}{\partial d_{ji}} \right) (\boldsymbol{z}_j - \boldsymbol{z}_i) \tag{56}$$

$$= -\sum_j \left[\frac{\tilde{p}(\boldsymbol{x}_j)}{\sum\limits_{j'} \tilde{p}(\boldsymbol{x}_{j'})\tilde{q}(\boldsymbol{z}_{j'})} - \frac{\tilde{q}(\boldsymbol{z}_j)}{\sum\limits_{j'} \tilde{q}^2(\boldsymbol{z}_{j'})} \right] 2 \frac{\partial \tilde{q}(\boldsymbol{z}_j)}{\partial d_{ij}} (\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{57}$$

Note that $\frac{\partial \tilde{q}(\boldsymbol{z}_j)}{\partial d_{ij}} = \frac{\partial \kappa_z(\boldsymbol{z}_j - \boldsymbol{z}_i)}{\partial d_{ij}} = 2\kappa^2(\boldsymbol{z}_j - \boldsymbol{z}_i))$ for $\kappa_z(\boldsymbol{z}_j - \boldsymbol{z}_i)) = \left[\frac{1}{1 + ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2} \right]$, which is the t-distribution, or alternatively for $\kappa_z(\boldsymbol{z}_j - \boldsymbol{z}_i)) = \exp(-\kappa||\boldsymbol{z}_i - \boldsymbol{z}_j||^2)$, the Gaussian distribution. Taken together,

$$\frac{\partial}{\partial \boldsymbol{z}_i} \widehat{CS}(p(\boldsymbol{x})||q(\boldsymbol{z})) = -4 \sum_j \left[\frac{\tilde{p}_j}{\sum\limits_{j'} \tilde{p}_{j'}\tilde{q}_{j'}} - \frac{\tilde{q}_j}{\sum\limits_{j'} \tilde{q}^2_{j'}} \right] \kappa_z^2(\boldsymbol{z}_j - \boldsymbol{z}_i)(\boldsymbol{z}_j - \boldsymbol{z}_i). \tag{58}$$

□

Proposition 9. *[Cauchy-Schwarz (CS) t-SNE is a special case of MAP IT]. Let $p_{jk'}$ be the probability for the joint event $\boldsymbol{x}_j \cap \boldsymbol{x}_{k'}$. Let $q_{jk'}$ be the probability for the joint event $\boldsymbol{z}_j \cap \boldsymbol{z}_k$. If $(\boldsymbol{x}_j \cap \boldsymbol{x}_{k'}) \cap (\boldsymbol{z}_j \cap \boldsymbol{z}_k) \in \emptyset$, then*

$$CS(P_m||Q_m) = CS(P||Q). \tag{59}$$

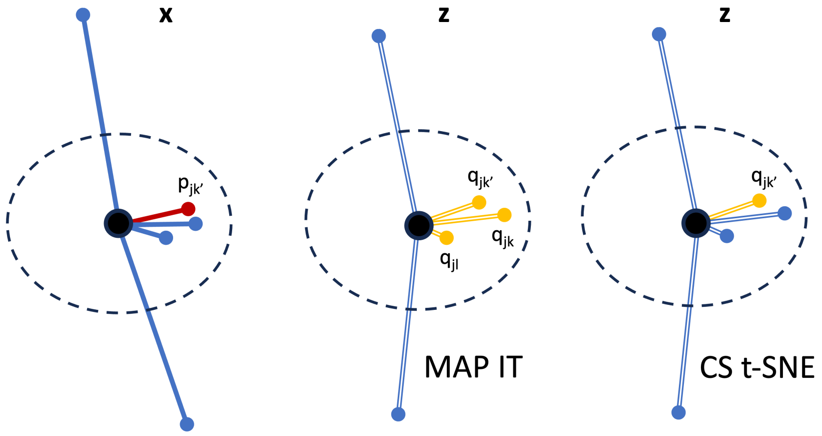


Figure 1: Illustration of CS t-SNE as a special case of MAP IT.

Proof. We have

$$CS(P_m||Q_m) = -\log \frac{\sum_j p_j q_j}{\left(\sum_j p_j^2\right)^{\frac{1}{2}} \left(\sum_j q_j^2\right)^{\frac{1}{2}}} \tag{60}$$

$$= -\log \frac{\sum_j \sum_{k',k} p_{jk'} q_{jk}}{\left(\sum_j \sum_{k',k} p_{jk'} p_{jk}\right)^{\frac{1}{2}} \left(\sum_j \sum_{k',k} q_{jk'} q_{jk}\right)^{\frac{1}{2}}}. \tag{61}$$

It suffices to look at the numerator since the terms in the denominator are related normalization quantities.

If $(\boldsymbol{x}_j \cap \boldsymbol{x}_{k'}) \cap (\boldsymbol{z}_j \cap \boldsymbol{z}_k) \in \emptyset$ then $Prob((\boldsymbol{x}_j \cap \boldsymbol{x}_{k'}) \cap (\boldsymbol{z}_j \cap \boldsymbol{z}_k)) = p_{jk'} q_{jk} = 0$ (assuming independence) for $k' \neq k$. Hence

$$\sum_j \sum_{k',k} p_{jk'} q_{jk} = \sum_{j,i} p_{ji} q_{ji} \tag{62}$$

for $k' = k = i$. □

Comment to Proposition 9. The illustration in Fig. 1 brings further perspective to this result.

The black filled circle denotes node j. Nodes within the stapled circle are assumed to be relatively near and nodes n outside this circle are assumed to be distant in the sense that p_{jn} is neglible for each such node n. For MAP IT, $p_{jk'}$ is also multiplied by probabilities q_{jk} and q_{jl} in addition to $q_{jk'}$ for nodes k and l close to k'. This is not the case for CS t-SNE.

In practise, this means that MAP IT models that the event $(\boldsymbol{x}_j \cap \boldsymbol{x}_{k'})$ could induce the event $(\boldsymbol{z}_j \cap \boldsymbol{z}_k)$ if k' and k are close.

3 Additional results and analysis

MNIST. MNIST[2] is a data set of 28×28 pixel grayscale images of handwritten digits. There are 10 digit classes (0 through 9) and a total of 70000 images. Here, 2000 images are randomly sampled. Each image is represented by a 784-dimensional vector. Figure 1 and Figure 3 in the main paper show that MAP IT produces a MNIST visualization with much better separation between classes compared to alternatives and that the embedding is robust with respect to initial conditions.

MAP IT's free parameter is the number of nearest neighbors k to go into the computation of $\tilde{p}_{j_{N_{\boldsymbol{x}_i}}}$ and $\tilde{q}_{j_{N_{\boldsymbol{x}_i}}}$. Figure 2 in this Supplementary shows representative embedding results for the subset of MNIST for different values of k. As in all dimensionality reduction methods, the visualization results depend on k. For MNIST, for $k = 7$ and for $k = 10$, the class structure appears. For $k = 12$ and up classes seem to be much compressed. For most data sets, a value for k between 5 and 15 seem to yield reasonable results. However, the impact of this hyperparameter should be further studied in future work.

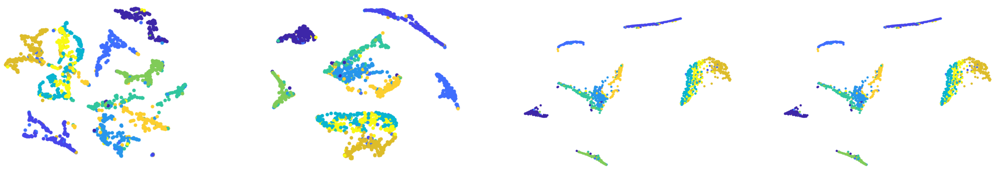
Figure 8 in the main paper shows the result of an initial experiment to scale up MAP IT by a certain sampling procedure. The proposed sampling procedure is not the same as the one employed in LargeVis and UMAP. In those methods, attractive forces and repulsive forces are separated. The number of points to go into the computation of repulsive forces are then sampled, so-called negative sampling. When creating Figure 8 in the main paper, forces have been separated into attractive/repulsive forces resulting from nearest neighbors versus attractive/repulsive forces coming from non-neighbors. Hence, the proposed MAP IT sampling is different. Experimentally, it was observed that if the number of non-neighbor forces were downsampled for instance to 50 percent, then a multiplication of the attractive/repulsive forces for non-neighbors by a factor two basically reproduced the original embedding. A downsampling of non-neighbor forces to 25 percent followed by a multiplication factor of four reproduced the original embedding. Similarly, downsampling of non-neighbor forces to 12.5 percent followed by a multiplication factor of eight reproduced the original embedding. This is illustrated in Figure 3 for MNIST with $k = 10$. In (a)-(c), the sampling is down to 50, 25, and 12.5 percent, respectively, and each of the four subfigures show the embedding after invoking a multiplication factor of 1, 2, 4 and 8 over the non-neighbor forces, respectively. The boxes indicates that the original embedding is in essence recreated (compare e.g. to Figure 2 (c)). For Figure 8 in the main paper, where only $3k$ non-neighbor forces are sampled, which means that ca 1.52 percent of non-neighbors are used in the sampling, the factor used is 66 (ca $1/0.0152$).

Learning rates (USPS). In all experiments the MAP IT learning rate has been set to 50 over 1000 iterations. Of course, changing these choices will to some degree change the embedding. These choices have however been observed to result in quite stable MAP IT results over a range of diverse data sets. Figure 4 shows the MAP IT cost as a function of iterations for different values of learning rates performed over a subset of the USPS data set (Hull, 1994). A random subset of the digits 3, 6, and 9 constitute the classes. For each learning rate η of 25, 50 and 100 three runs of MAP IT are performed and in each case the curve of cost versus iterations is shown. For this particular data set, low cost function values are obtained quicker for $\eta = 100$ (leftmost group of curves), compared to $\eta = 50$ (middle group of curves) and $\eta = 25$ (rightmost group of curves). When approaching 1000 iterations all curves have settled at low cost function values. Further studies of the interplay between learning rate, iterations, and various design choices for the MAP IT optimization are left for future work.

Coil 20. This data set (Nene et al., 1996) consists of 1440 greyscale images consisting of 20 objects under 72 different rotations spanning 360 degrees. Each image is a 128x128 image which we treat as a single 16384 dimensional vector for the purposes of computing distance between images. Visualizations of Coil-20 were shown in the main paper in Figure 4. Enlarged visualizations of Coil 20 are shown in Figures 5, 6, 7 and 8.

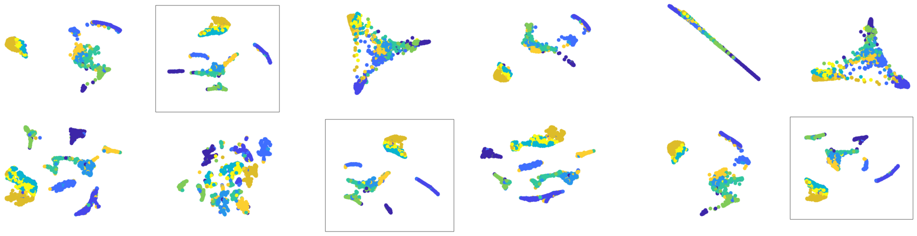
Visual Concepts. Images corresponding to three different visual concepts are visualized. SIFT (Lowe, 1999) descriptors represented by a 1000-dimensional codebook for each visual concept are downloaded from the ImageNet data base (`image-net.org`) (Deng et al., 2009). The visual concepts used are *strawberry*, *lemon* and *australian terrier*. The concepts are represented by 1478, 1292 and 1079 images, respectively. The images within each category differ very much, as can be seen

[2] MNIST, Newsgroups, "Frey faces" are obtained from `http://cs.nyu.edu/~roweis/data.html`.



(a) $k = 5$. (b) $k = 7$. (c) $k = 10$. (d) $k = 12$?.

Figure 2: MAP IT for a subset of MNIST for different values of k.



(a) 50 percent downsampling. (b) 25 percent downsampling. (c) 12.5 percent downsampling.

Figure 3: Each subfigure (a)-(c) show the visualization/embedding for different subsampling scenarios and for a factor 1, 2, 4, and 8, respectively, on non-neighbor attractive/repulsive forces in the MAP IT calculation.

e.g. for *australian terrier* at `image-net.org/synset?wnid=n02096294`. A crude approach is taken here. Each image is represented by the overall frequency of codewords present for the SIFT descriptors contained in the image. Hence, each image is represented as a 1000-dimensional vector. The local modeling strength of the SIFT descriptors are lost this way, and one cannot expect the resulting data set to contain very discriminative features between the concepts. Visualizations of the visual concepts were shown in the main paper in Figure 5. Enlarged visualizations of Coil 20 are shown in Figures 9, 10, 11 and 12.

Newsgroups. Visualizations of words from Newsgroups were shown in the main paper in Figure 6. Enlarged visualizations of Newsgroups as word clouds are shown in Figures 13, 14, 15 and 16.

Frey faces. Visualizations of the Frey faces were shown in the main paper in Figure 7. Enlarged visualizations of the Frey faces are shown in Figures 17, 18, 19 and 20.

Together with the experiments and analysis in the main paper, these additional MAP IT results and analysis illustrate the potential of this new method to provide visualizations which in many cases are markedly different from the current state-of-the-art alternative with better class discrimination and reasonable embeddings overall, from a theoretical approach which is fundamentally different and which highlights both a viewpoint from the perspective of alignment of marginal probabilities as well as a dual viewpoint via continuous densities enabled by kernel smoothing. The role of normalization follows directly from the theory.

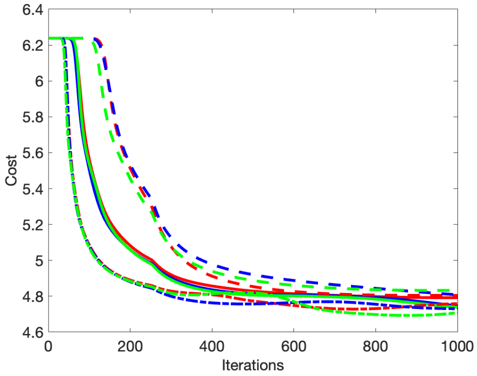


Figure 4: Illustration of MAP IT learning rates and number of iterations for a subset of USPS.

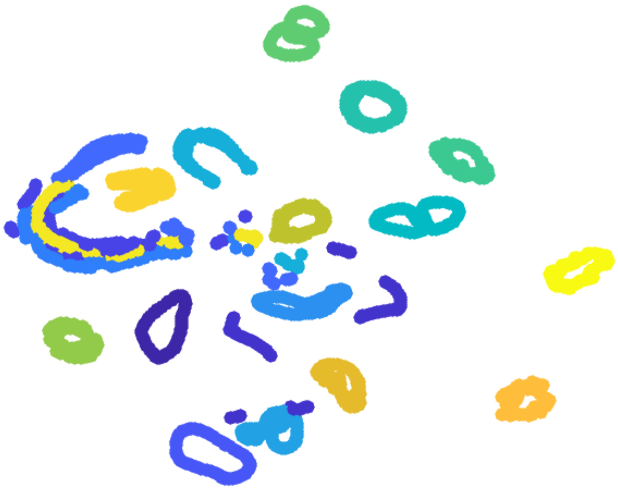


Figure 5: t-SNE embedding of Coil 20.

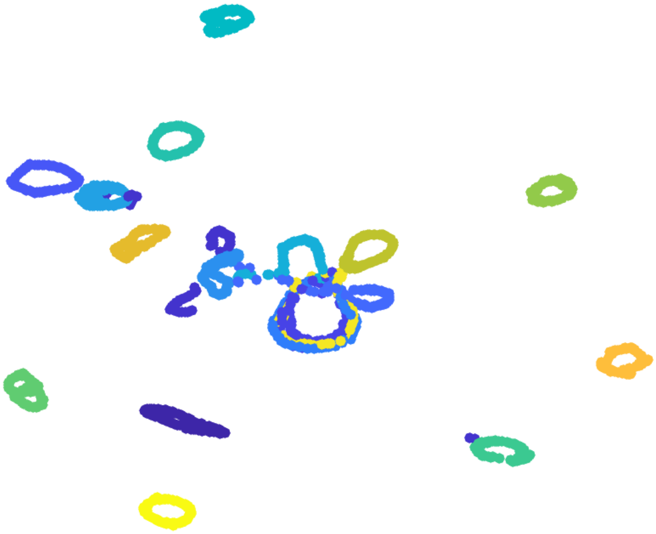


Figure 6: UMAP embedding of Coil 20.

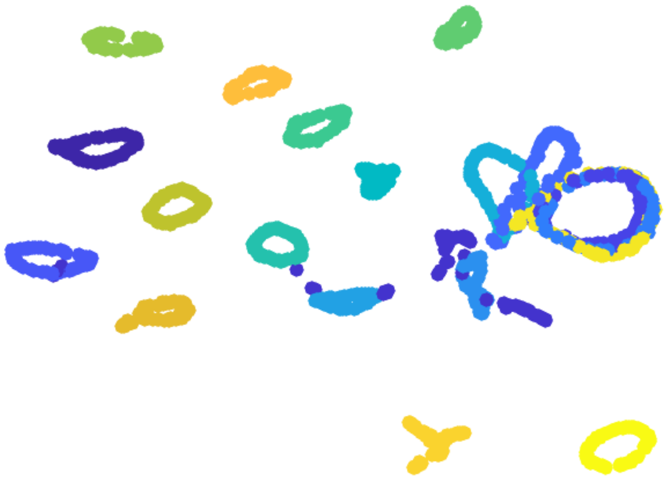


Figure 7: PacMap embedding of Coil 20.

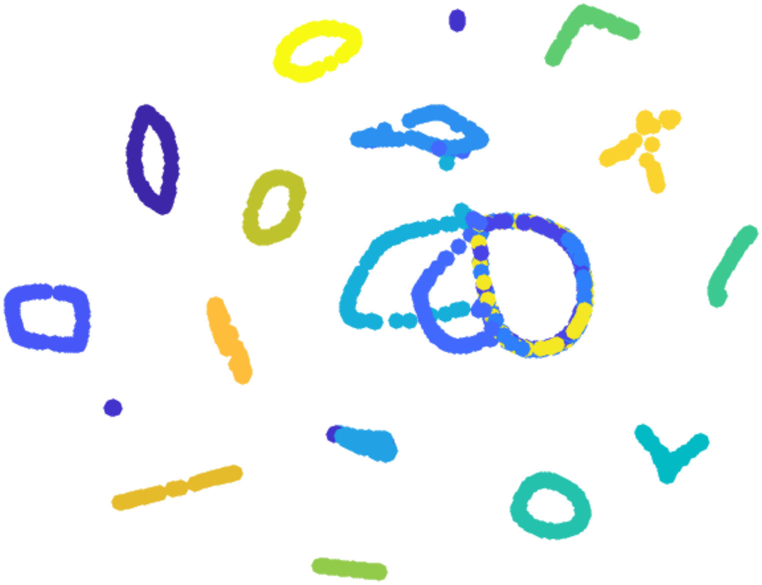


Figure 8: MAP IT embedding of Coil 20.

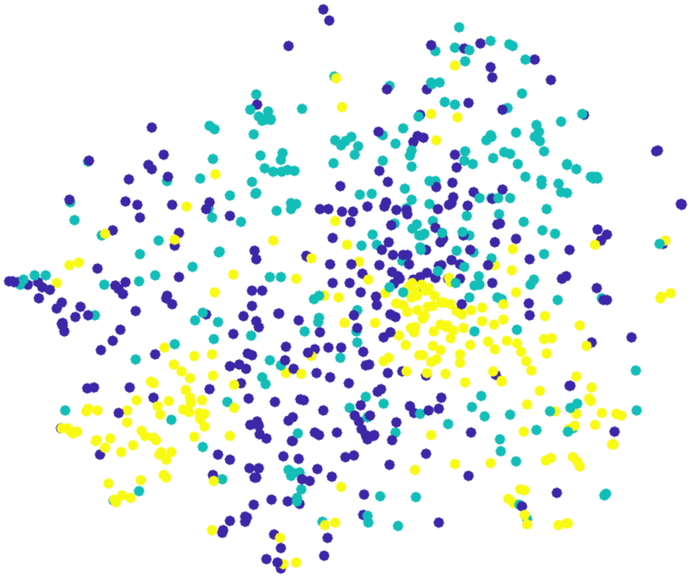


Figure 9: t-SNE embedding of visual concepts.

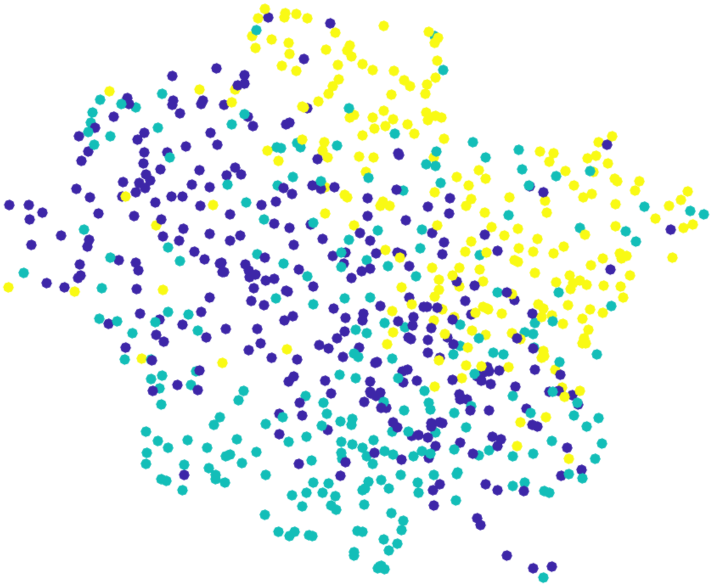


Figure 10: UMAP embedding of visual concepts..

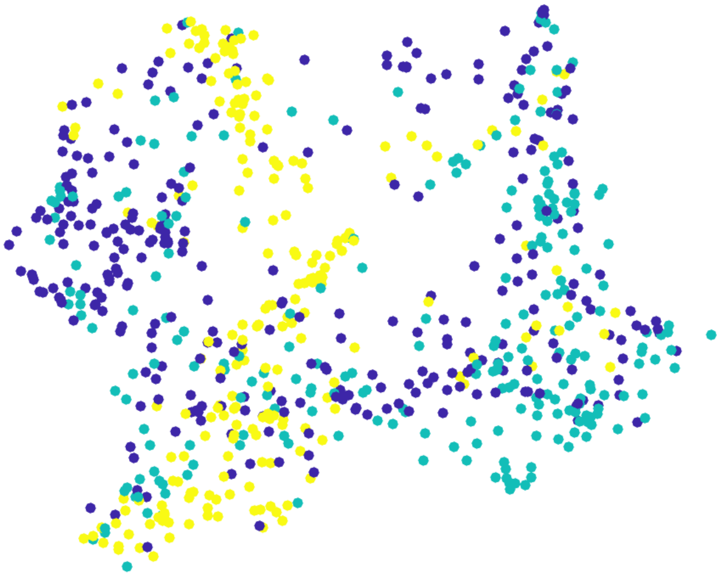


Figure 11: PacMap embedding of visual concepts..

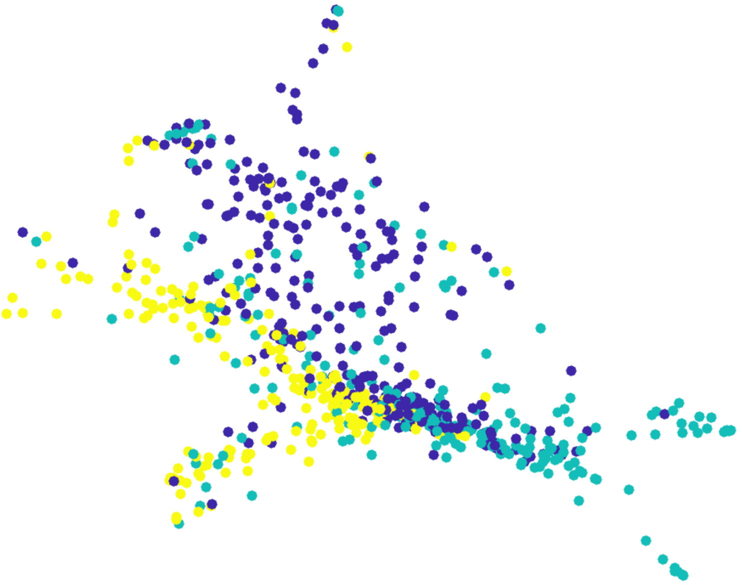


Figure 12: MAP IT embedding of visual concepts..

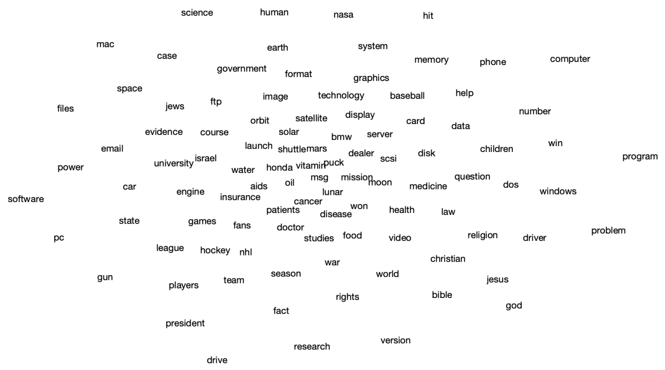


Figure 13: t-SNE embedding of words from Newsgroups.

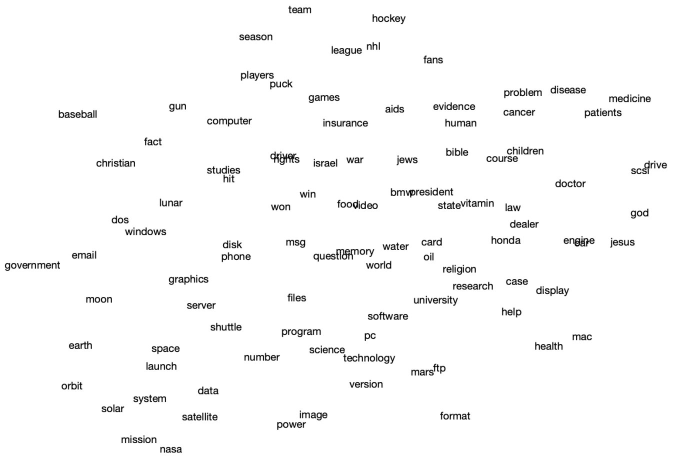


Figure 14: UMAP embedding of words from Newsgroups.

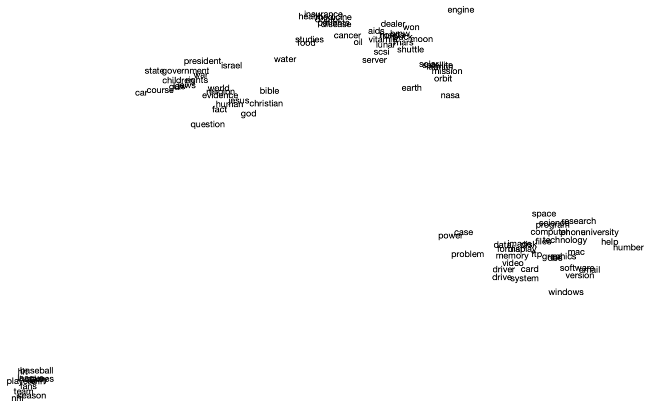


Figure 15: PacMap embedding of words from Newsgroups.

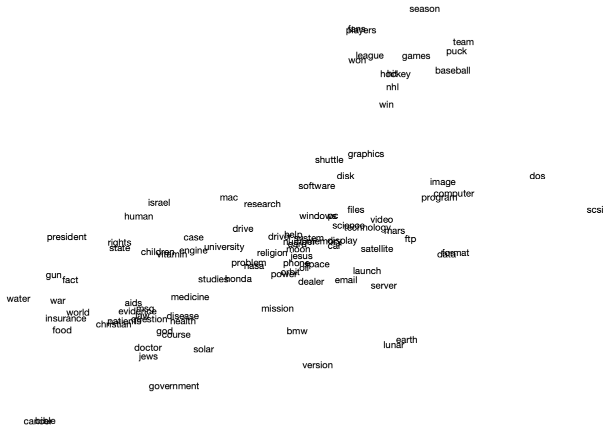


Figure 16: MAP IT embedding of words from Newsgroups..

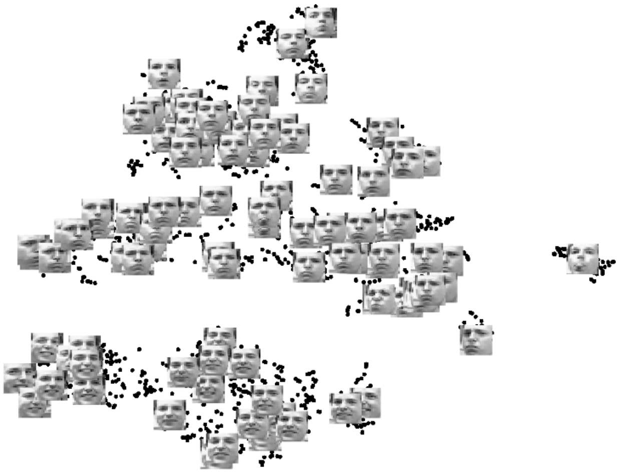


Figure 17: t-SNE embedding of Frey faces.

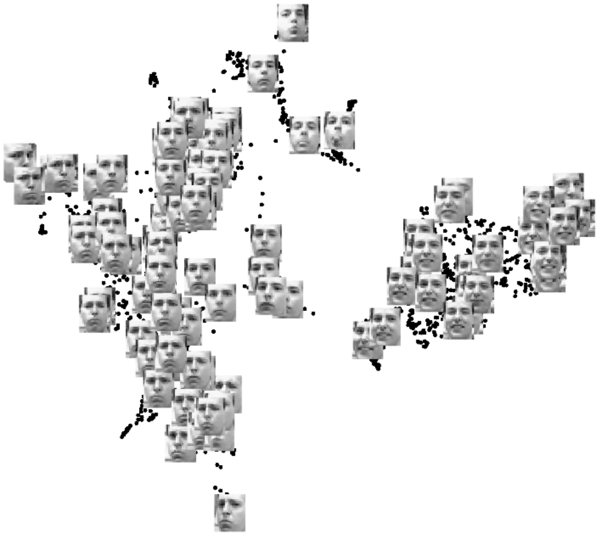


Figure 18: UMAP embedding of Frey faces.

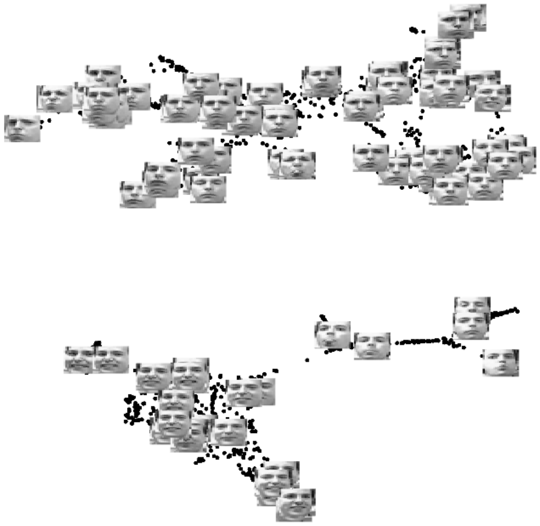


Figure 19: PacMap embedding of Frey faces.

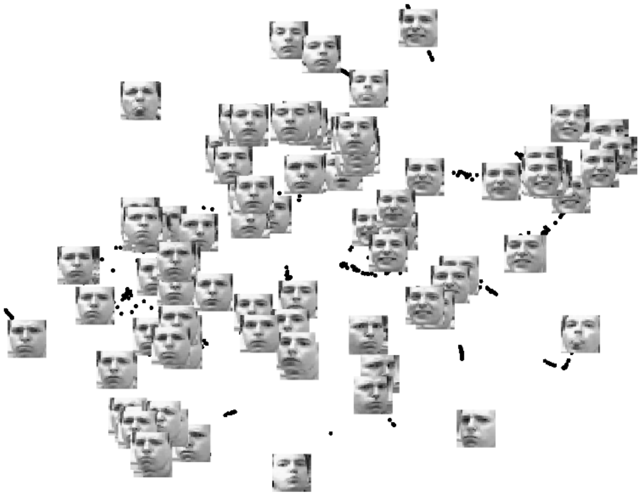


Figure 20: MAP IT embedding of Frey faces.

Acknowledgments

There will be some acknowledgments, including those to funding agencies. Will probably span two/three sentences. This is to make some space for that.

References

Ehsan Amid and Manfred K Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

Jan Niklas Bohm, Philipp Berens, and Dmitry Kobak. A unifying perspective on neighbor embeddings along the attraction-repulsion spectrum. *arXiv preprint arXiv:2007.08902*, 2020.

Kerstin Bunte, Sven Haase, and Thomas Villmann. Stochastic neighbor embedding SNE for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

T. Tony Cai and Rong Ma. Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 23(1):1–54, 2022.

Sebastian Damrich and Fred A. Hamprecht. On UMAP's true loss function. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

Andrew Draganov, Jacob Rødsand Jørgensen, Katrine Scheel Nelleman, Davide Mottin, Ira Assent, Tyrus Berry, and Aslay Cigdem. ActUp: Analyzing and consolidating t-SNE and UMAP. *arXiv preprint arXiv:2305.07320*, 2023.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pp. 1735–1742, 2006.

Geoffrey E. Hinton and Sam Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

Yanyong Huang, Kejun Guo, Xiuwen Yi, Jing Yu, Zongxin Shen, and Tianrui Li. T-copula and wasserstein distance-based stochastic neighbor embedding. *Knowledge-based Systems*, 243: 108431, 2022.

J. J. Hull. A Database for Handwritten Text Recognition Research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021.

David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*. IEEE, 1999.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2020.

Karthik Naryan, Ali Punjani, and Pieter Abbeel. Alpha-beta divergences discover micro and macro structures in data. In *International Conference on Machine Learning*, volume 37. 2015.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20). In *Technical Report*. 1996.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *International Conference on World Wide Web*, volume 25. ACM, 2016.

Daniel J. Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Robert Jenssen, and Michael Kampffmeyer. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7527–7536. IEEE, 2023.

Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3321–3245, 2014.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Laurens van der Maaten and Killian Weinberger. Stochastic triplet embedding. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2012.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PacMAP for data visualization. *Journal of Machine Learning Research*, 22(1):9129–9201, 2021.

Michael Wilber, Iljung Kwak, David Kriegman, and Serge Belongie. Learning concept embeddings with combined human-machine expertise. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, 2015.