Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Journal Track

Improving the Effectiveness and Efficiency of Stochastic Neighbour Embedding With Isolation Kernel (Extended Abstract) *

Ye Zhu[1] , **Kai Ming Ting**[2]
[1]School of Information Technology, Deakin University, Geelong, Australia
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Jiangsu, China
ye.zhu@ieee.org, tingkm@nju.edu.cn

Abstract

This paper presents a new insight into improving the performance of Stochastic Neighbour Embedding (t-SNE) by using Isolation kernel instead of Gaussian kernel. We show that Isolation kernel addresses two deficiencies of t-SNE that employs Gaussian kernel, and the use of Isolation kernel enables t-SNE to deal with large-scale datasets in less runtime without trading off accuracy, unlike existing methods used in speeding up t-SNE.

1 Introduction and Motivation

t-SNE [Maaten and Hinton, 2008] has been a successful and popular dimensionality reduction method for visualisation. It aims to project high-dimensional datasets into lower-dimensional spaces while preserving the similarities between data points, as measured by the KL divergence.

The original SNE [Hinton and Roweis, 2003] employs a Gaussian kernel to measure similarity in both high and low-dimensional spaces. t-SNE replaces the Gaussian kernel with the distance-based similarity $(1 + d_{ij})^2$ (where d_{ij} is the distance between instances i and j) in low-dimensional space, while retaining the Gaussian kernel for high-dimensional space.

When using the Gaussian kernel, t-SNE has to fine-tune a bandwidth of the Gaussian kernel centred at each point in a given dataset because Gaussian kernel is independent of data distribution. In other words, t-SNE must determine n bandwidths for a dataset of n points.

The bandwidth determination process in t-SNE uses a heuristic search with a single global parameter called perplexity such that the Shannon entropy is fixed for all probability distributions at all points while adapting each bandwidth to the local density of the dataset. As the perplexity can be interpreted as a smooth measure of the effective number of neighbours [Maaten and Hinton, 2008], the method can be interpreted as using a user-specified number of nearest neighbours (aka kNN) in order to determine the n bandwidths. Whilst there is a single external parameter $perplexity$, a bandwidth setting must be optimised for each data point internally.

This becomes the first obstacle in dealing with large datasets due to the massive computational cost of the bandwidth search process. In addition, the point-based bandwidth is also the cause of misrepresentation in high-dimensional space under some conditions.

To date, the common practice is still using Gaussian kernel in t-SNE. Yet, sound and workable solutions to its drawbacks mentioned above are not available.

The contributions of this paper are:

(1) Uncovering two deficiencies due to the use of Gaussian kernel. First, the point-based-bandwidth Gaussian kernel often creates misrepresented structure(s) that do not exist in high-dimensional space under some conditions. Second, the use of the data-independent kernel requires t-SNE to determine n bandwidths for a dataset of n points, despite the fact that a user needs to set one parameter only. This becomes one key obstacle in dealing with large datasets.

(2) Revealing the advantages of using a partition-based data-dependent kernel in t-SNE. First, this kernel enables the true structure(s) in the high-dimensional space to be represented correctly under the same condition mentioned above. Second, the data-dependent similarity is set with a single parameter only; this allows it to be computed more efficiently. These two advantages enable t-SNE to deal with large-scale datasets without trading off accuracy with faster runtime, without resorting to approximation methods.

(3) Proposing an improvement to t-SNE by simply replacing the data-independent kernel with a data-dependent kernel, leaving the rest of the procedure unchanged.

(4) Verifying the effectiveness and efficiency of the data-dependent kernel in t-SNE.

While many improved methods have been proposed since the introduction of SNE (e.g., [Cook *et al.*, 2007; Yang *et al.*, 2009; Venna *et al.*, 2010; Van Der Maaten and Weinberger, 2012; Lee *et al.*, 2013; Van Der Maaten, 2014; Linderman and Steinerberger, 2017; Shaham and Steinerberger, 2017; Arora *et al.*, 2018; Linderman *et al.*, 2019]), none of them has investigated the suitability of Gaussian kernel in t-SNE.

Since t-SNE needs a data-dependent kernel, we propose to use a recent data-dependent kernel called Isolation kernel [Ting *et al.*, 2018; Qin *et al.*, 2019] to replace the data-independent Gaussian kernel in t-SNE. The Isolation kernel is

*This work is originally published in the Journal of Artificial Intelligence Research, 71:667–695, August 2021.

a perfect match for the task because a data-dependent kernel, by definition, adapts to local distribution without any additional optimisation. The kernel replacement is conducted in the component in the high-dimensional space only, leaving the other components of the t-SNE procedure unchanged.

The experiment result shows that using Isolation kernel improves the performance of t-SNE and resolves the issues brought by Gaussian kernel in t-SNE.

2 Two Deficiencies of Gaussian Kernel When Used in t-SNE

2.1 The First Deficiency

Gaussian kernel determines each local bandwidth based on one local point only. It often creates misrepresented structure(s) that do not exist in high-dimensional space under some conditions.

Point-Based Bandwidth: The Cause of Misrepresentation in High-Dimensional Space

As bandwidth σ_i of the Gaussian kernel is fixed for each point x_i, we have the following observation:

Observation 1. *Gaussian kernel with point-based bandwidth can misrepresent the structure of a data distribution that has points significantly denser than the majority of the points in a sample generated from the distribution.*

An example simulation result of t-SNE's misrepresentation is shown in the first column in Figure 1: t-SNE is unable to identify the joint component of the three clusters in different subspaces which share the same mean at the origin only in the high-dimensional space but nowhere else. Notice that the mapped origin point is misrepresented to be associated with one cluster only; and it is totally disassociated with the other two clusters.[1]

2.2 The Second Deficiency

The use of the data-independent kernel requires t-SNE to determine n bandwidths for a dataset of n points. This becomes one key obstacle in dealing with large datasets.

Low Computational Efficiency With Gaussian Kernel

The use of a Gaussian kernel necessitates the search for a local bandwidth for each local point. t-SNE utilises a binary search for the value of σ_i that makes the entropy of the distribution over neighbours equal to $\log K$, where K is the effective number of local neighbours or "perplexity" [Maaten and Hinton, 2008]. This search is the key component that determines the success or failure of t-SNE. A gradient descent search has been used successfully to perform the search for n parameters for small datasets [Maaten and Hinton, 2008]. This formulation has two key limitations for large datasets. First, the need for n-parameters search poses a real limitation in terms of finding appropriate settings for a large number of parameters. Second, it cannot deal with large datasets because it has high time complexity $O(n^2)$.

[1]A Matlab code demonstration of using t-SNE with Isolation kernel can be obtained from https://github.com/zhuye88/IKt-sne.

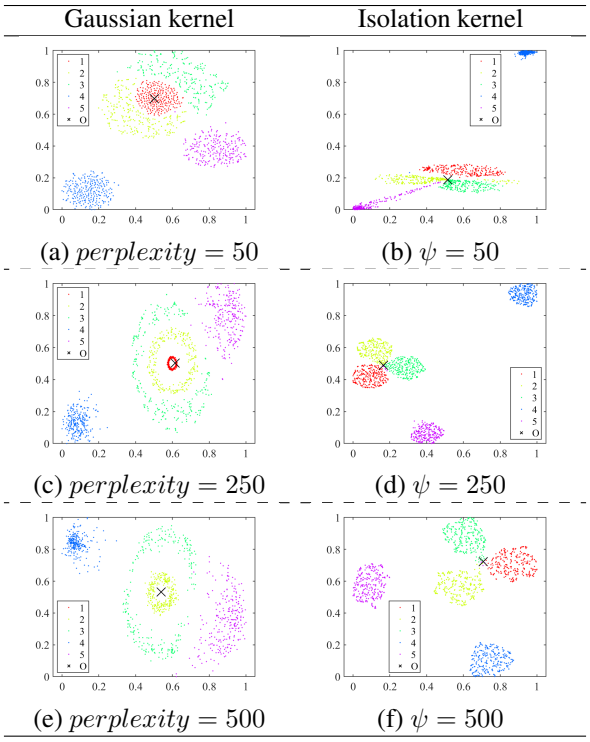


Figure 1: Visualisation results of t-SNE using Gaussian kernel or Isolation kernel on a 50-dimensional dataset with 5 subspace clusters, each in a different 10-dimensional subspace. The black cross indicates the mapped point of the origin in the high-dimensional space shared by three clusters in different subspaces. Note that in (c), all points of the red cluster (cluster 1) are concentrated and they overlap with the mapped origin. $perplexity$ and ψ are the key parameters for Gaussian kernel and Isolation kernel, respectively.

3 Using the Isolation Kernel in t-SNE

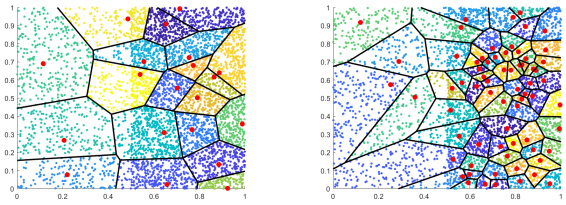
Since t-SNE needs a data-dependent kernel, we propose to use a recent data-dependent kernel called Isolation kernel [Ting *et al.*, 2018; Qin *et al.*, 2019] to replace the data-independent Gaussian kernel in t-SNE.

The Isolation kernel is a perfect match for the task because a data-dependent kernel, by definition, adapts to local distribution without any additional optimisation. The kernel replacement is conducted in the component in the high-dimensional space only, leaving the other components of the t-SNE procedure unchanged.

3.1 Isolation Kernel Versus Gaussian Kernel

The key difference is that the Isolation kernel adapts to local density distribution, but the Gaussian kernel is independent of the data distribution.

In addition, the technical differences can be observed in two aspects. First, the Isolation kernel has no closed-form expression. Second, it is derived directly from a dataset, without explicit learning or optimisation. Its adaptation to local density is a direct outcome of its isolation mechanism used to partition space, i.e., the mechanism produces *large partitions in sparse regions and small partitions in dense regions* [Ting *et al.*, 2018; Qin *et al.*, 2019]. A natural isolation mechanism that has this



(a) $\psi = 16$ (b) $\psi = 64$

Figure 2: Two examples of partitioning H using the nearest neighbour (a Voronoi diagram) on a dataset having two regions of uniform densities, where the left half has a lower density than the right half.

characteristic is a Voronoi diagram. Given a sample of the underlining distribution, each Voronoi cell isolates a point from the rest of the points in the sample; and the cells are small in the dense region and large in the sparse region.

Note that the Voronoi diagram is obtained very efficiently, i.e., given a sample, nothing else needs to be done in the training stage because boundaries in the Voronoi diagram can be obtained at the testing stage as the equal distance between the two nearest points in the given sample.

Figure 2 shows two examples of partitioning H using the nearest neighbour or a Voronoi diagram on the same dataset with two different subsample sizes ψ. These examples show that there are more (small) cells in the dense region than (large) cells in the sparse region for each ψ; and the sizes of the cells are usually decreasing with respect to ψ. Two points located in the same cell get a similarity score of 1 in a partitioning. The final Isolation kernel similarity between two points is the probability of both points falling into the same cell over a finite number of partitionings.

The Isolation kernel has the following well-defined data-dependent characteristic: **two points in a sparse region are more similar than two points of equal inter-point distance in a dense region** [Ting *et al.*, 2018]. The details of Isolation kernel are provided in [Ting *et al.*, 2018; Qin *et al.*, 2019].

3.2 The Isolation Kernel Makes Full Use of the Distributional Information in Small Samples

The Isolation kernel only requires small samples (ψ) for the space partitioning without a computationally expensive process. A small sample of a dataset contains data distributional information which is sufficient to build a data-dependent kernel. The Isolation kernel extracts this information in the form of a Voronoi diagram, which depicts the relative densities between regions.

In contrast, using a data-independent measure such as the Gaussian kernel, the distributional information in a dataset is ignored and each point in the input space is treated as an independent point. In order to get the distributional information in the form of variable bandwidths that are adaptive to the local distribution, a separate optimisation process is required, as conducted in step 1 of the t-SNE algorithm.

It is important to note that when they could not handle a large dataset, most methods often use a small sample as a mitigation approach, and this inevitably trades off runtime with accuracy. But it is not the case for the Isolation kernel where small samples are the key to achieving high accuracy[2]; and sample sizes larger than the optimal ψ degrade the accuracy of Isolation kernel.

	Gaussian kernel	Isolation kernel
Step 1: Kernel building		$\mathcal{O}(t\psi)$
Bandwidth search	$\mathcal{O}(rn^2)$	
Step 2: Matrix calculation	$\mathcal{O}(m^2)$	$\mathcal{O}(t\psi m^2)$
Step 3: t-SNE Mapping	$\mathcal{O}(sm^2)$	

Table 1: Time complexities of t-SNE in steps (1) kernel building (IK) or Bandwidth search (GK), (2) computing the similarity, and (3) mapping from high to low dimensions. r is the number of iterations used for bandwidth search for the Gaussian kernel; and s is the number of iterations in t-SNE mapping. $m(\leq n)$ is the subsample size used for the mapping. For small datasets: $m = n$.

In other words, by using the Gaussian kernel, t-SNE must employ a computationally expensive approach to get the distributional information in a dataset. It does not exploit the same information which is freely available in small samples of the dataset. The Isolation kernel is a direct approach that makes full use of the distributional information freely available in small samples of a dataset.

3.3 Overcoming the Two Deficiencies

The Isolation Kernel's Parameter Is Not Point-Based

Isolation kernel with one parameter ψ ensures that the local structure is truly reflected in the similarities among local points in the high-dimensional space: the three clusters are well separated and yet they share a common point, indicated by the mapped origin point as shown in the second column in Figure 1, unlike the misrepresentation exhibited in the first column when the Gaussian kernel is used.

This is possible because Isolation kernel is partition-based, points in the local neighbourhood are most likely to be in the same partition. As a result, t-SNE using the Isolation kernel produces an improved visualisation quality that has no misrepresentations.

High Computational Efficiency With Isolation Kernel

The computational complexities of the Gaussian kernel and Isolation kernel [Ting *et al.*, 2018; Qin *et al.*, 2019] used in t-SNE are shown in Table 1. Although the parameter ψ of

[2]Isolation kernel performs optimally with small samples was formally analysed in the context of nearest neighbour anomaly detection [Ting *et al.*, 2017]. The work is motivated by the previous finding that small samples can produce better detection accuracy for some anomaly detectors than large samples (e.g., [Liu *et al.*, 2008; Sugiyama and Borgwardt, 2013].) The theoretical analysis based on computational geometry reveals that the geometry of data distribution has a direct impact on the sample size setting which is essential to produce an optimal nearest neighbour anomaly detector [Ting *et al.*, 2017]. In a simple geometry such as a Gaussian distribution, a sample size of one data point (at the mean of Gaussian distribution) yields the optimal nearest neighbour anomaly detector; and a sample of more data points will produce a worse performing detector. In a more complex geometry of data distribution (e.g., a mixture of multiple Gaussian distributions), while the optimal sample size is more than one data point, a sample size more than the optimal one also produces a worse performing detector. See [Ting *et al.*, 2017] for details.

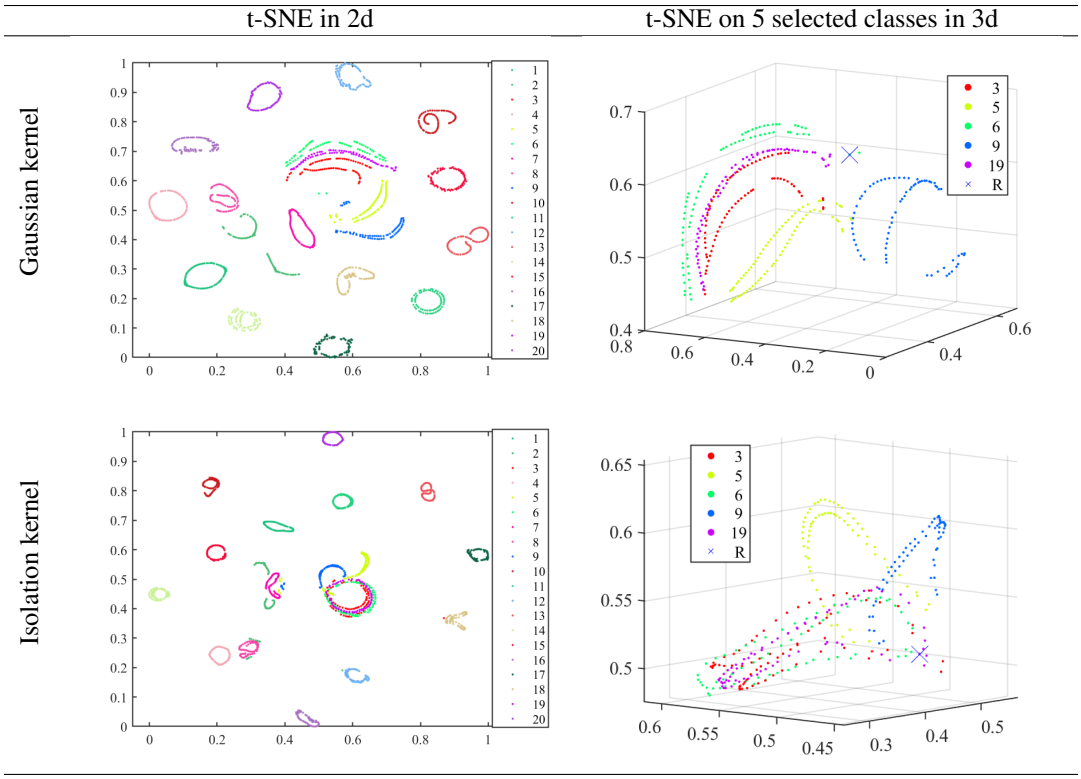


Figure 3: The first column shows the t-SNE visualisation results on COIL20 in a two-dimensional space. The second column shows the five clusters and a reference point (indicated as $\times$ with the class label "R") on t-SNE visualisation results in a three-dimensional space.

Isolation kernel corresponds to the bandwidth parameter of the Gaussian kernel, the Isolation kernel needs no optimisation to determine n bandwidths locally. This is because the partitioning mechanism used by the Isolation kernel produces small partitions in dense regions and large partitions in sparse regions; and the sizes of the partitions are monotonically decreasing with respect to ψ. As the local adaptation has already been done during the process of deriving the kernel, no further adaptation is required after the kernel is derived.

Though the Isolation kernel derived from a dataset takes constant $O(t\psi)$ time, it is significantly less than the optimisation required to determine n bandwidths which takes $O(n^2)$ time in Gaussian kernel. For a large dataset, when using Gaussian kernel, it is infeasible to estimate a large number of bandwidths with an appropriate degree of accuracy, and its computational cost is prohibitively high.

The consequence of using Isolation kernel is that the runtime of step 1 in the t-SNE algorithm is significantly reduced. This enables the Isolation kernel version of t-SNE to deal with large datasets.

4 Example Evaluations

Here we provide two example evaluations. First, on a real-world dataset COIL20, we have identified a structural misrepresentation issue with the Gaussian kernel, similar to the one shown in Figure 1.

Figure 3 shows the five clusters where the Gaussian kernel has misrepresented structures in the high-dimensional space. The 3-dimensional results denote that the Isolation kernel depicts a more nuanced structural relationship between the five clusters; whereas the Gaussian kernel depicts that they are disparate five clusters, shown in the second column in Figure 3. Also, note that a reference point $\times$ is close to all five clusters when the Isolation kernel is used, but it is far from many clusters when a Gaussian kernel is used.

Second, though both Gaussian Kernel and Isolation kernel have quadratic time and space complexities, using Gaussian kernel in t-SNE needs a large number of iterations to search for the optimal local bandwidth for each point. As a result, t-SNE takes a much longer time in step 1 of the algorithm than the Isolation kernel version of t-SNE. For example, in a dataset of 100,000 data points, the Isolation kernel version of t-SNE is two orders of magnitude faster than the Gaussian kernel version (887 seconds versus 72,196 seconds).

Further details of the proposed method and the evaluation results can be found in the full paper [Zhu and Ting, 2021]. This includes existing methods FIt-SNE [Linderman *et al.*, 2019] and opt-SNE [Belkina *et al.*, 2019] that employ an approximation approach to address the efficiency issue of t-SNE, but the misrepresentation issue remains because they all use Gaussian kernel.

In a nutshell, the proposed method of using Isolation kernel is a unique approach to address the two deficiencies of t-SNE from their root cause that no existing methods can provide.

References

[Arora *et al.*, 2018] Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. *arXiv preprint arXiv:1803.01768*, 2018.

[Belkina *et al.*, 2019] Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):1–12, 2019.

[Cook *et al.*, 2007] James Cook, Ilya Sutskever, Andriy Mnih, and Geoffrey Hinton. Visualizing similarity data with a mixture of maps. In *Artificial Intelligence and Statistics*, pages 67–74, 2007.

[Hinton and Roweis, 2003] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

[Lee *et al.*, 2013] John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[Linderman and Steinerberger, 2017] George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. *arXiv preprint arXiv:1706.02582*, 2017.

[Linderman *et al.*, 2019] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

[Liu *et al.*, 2008] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*, pages 413–422, 2008.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Qin *et al.*, 2019] Xiaoyu Qin, Kai Ming Ting, Ye Zhu, and CS Vincent Lee. Nearest-neighbour-induced isolation similarity and its impact on density-based clustering. In *Thirty-third AAAI Conference on Artificial Intelligence*, 2019.

[Shaham and Steinerberger, 2017] Uri Shaham and Stefan Steinerberger. Stochastic neighbor embedding separates well-separated clusters. *arXiv preprint arXiv:1702.02670*, 2017.

[Sugiyama and Borgwardt, 2013] Mahito Sugiyama and Karsten Borgwardt. Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems 26*, pages 467–475, 2013.

[Ting *et al.*, 2017] Kai Ming Ting, Takashi Washio, Jonathan R. Wells, and Sunil Aryal. Defying the gravity of learning curve: a characteristic of nearest neighbour anomaly detectors. *Machine Learning*, 106:55–91, 2017.

[Ting *et al.*, 2018] Kai Ming Ting, Yue Zhu, and Zhi-Hua Zhou. Isolation kernel and its effect on SVM. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2329–2337. ACM, 2018.

[Van Der Maaten and Weinberger, 2012] Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

[Van Der Maaten, 2014] Laurens Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[Venna *et al.*, 2010] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb):451–490, 2010.

[Yang *et al.*, 2009] Zhirong Yang, Irwin King, Zenglin Xu, and Erkki Oja. Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 2169–2177, 2009.

[Zhu and Ting, 2021] Ye Zhu and Kai Ming Ting. Improving the effectiveness and efficiency of stochastic neighbour embedding with isolation kernel. *Journal of Artificial Intelligence Research*, 71:667–695, August 2021.