Energy Models for Better Pseudo-Labels: Improving Semi-Supervised Classification with the 1-Laplacian Graph Energy

Angelica I Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Samar M Alsaleh, Robby T Tan and Carola-Bibiane Schönlieb

Abstract—Semi-supervised classification is a great focus of interest, as in real-world scenarios obtaining labels is expensive, time-consuming and might require expert knowledge. This has motivated the fast development of semi-supervised techniques, whose performance is on a par with or better than supervised approaches. A current major challenge for semi-supervised techniques is how to better handle the network calibration and confirmation bias problems for improving performance. In this work, we argue that energy models are an effective alternative to such problems. With this motivation in mind, we propose a hybrid framework for semi-supervised classification called CREPE model (1-Laplacian Graph Energy for Pseudo-labelling). Firstly, we introduce a new energy model based on the non-smooth $\ell_1$ norm of the normalised graph 1-Laplacian. Our functional enforces a sufficiently smooth solution and strengthens the intrinsic relation between the labelled and unlabelled data. Secondly, we provide a theoretical analysis for our proposed scheme and show that the solution trajectory does converge to a non-constant steady point. Thirdly, we derive the connection of our energy model for pseudo-labelling. We show that our energy model produces more meaningful pseudo-labels than the ones generated directly by a deep network. We extensively evaluate our framework, through numerical and visual experiments, using six benchmarking datasets for natural and medical images. We demonstrate that our technique reports state-of-the-art results for semi-supervised classification.

Index Terms—Semi-Supervised Learning, Energy Models, Graph Laplacian, Pseudo-Labelling, Image Classification, Deep Learning

1 INTRODUCTION

In this era of big data, deep learning (DL) has reported astonishing results for different tasks in computer vision. For the task of image classification, a major breakthrough has been reported in the setting of supervised learning. In this context, the majority of methods are based on deep convolutional neural networks e.g. [1], [2], in which pre-trained, fine-tuned and trained from scratch solutions have been considered. A key factor for these impressive results is the assumption of a large and well-representative corpus of labelled data. These labels can be generated either by humans or automatically on proxy tasks. However, obtaining well-annotated labels is expensive, time consuming and one should account for the inherent human bias and uncertainty that adversely effect the classification output. These drawbacks have motivated semi-supervised learning (SSL) [3], [4] to be a great focus of interest for the community.

The key idea of SSL is to leverage on a tiny labelled set and a large unlabelled set to produce a good classification output. The desirable advantages of this setting is that one can decrease the dependency on a large amount of well-annotated data whilst gaining further understanding of intrinsic data structures [3]. The body of literature has reported promising results, from the classic perspective, for semi-supervised classification using both transductive (e.g [5], [6], [7], [8], [9], [10], [11]) and inductive (e.g. [12], [13], [14]) philosophies. Those techniques seek to infer the labels for the large unlabelled set, relying solely on the tiny labelled set as prior, by minimising a given energy (i.e., energy models [3], [5], [6]). More recently, deep learning has also been applied in the SSL context - examples are [15], [16], [17], [18], where strong augmentations and costly optimisation schemes are key for the outstanding performance. Both perspectives have shown potentials but they still have limitations. Whilst energy models rely on hand-crafted features that are difficult to generalise, deep learning techniques lack of a well-defined theory.

A few recent works [19], [20], [21] have attempted to combine both perspectives, so-called hybrid models, where the principles of energy models and deep learning are combined. Hybrid models have demonstrated performance which readily competes against deep learning techniques [21]. However, similarly to deep learning techniques, hybrid models have been investigated mainly from the practical point of view. That is, in the context of hybrid models not much effort has been spent on developing better energy functionals and analysing their theoretical properties. This is the motivation that drives the basis for this work.

More precisely, in this work we propose a robust graph energy for semi-supervised classification following a hybrid setting. We focus on the normalised Dirichlet energy (2) based on the graph Laplacian. Promising results have already been shown in this context. For example, the seminal

- Al Aviles-Rivero, P sellars and CB Schönlieb are with Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK (ai323, ps644, cbs31@cam.ac.uk)
- N Papadakis is with the IMB, Université Bordeaux, France. (nicolas.papadakis@math.u-bordeaux.fr)
- R Li is with ByteDance, Singapore. rhein@bytedance.com
- S Alsaleh is with the Department of Computer Science, Taibah University, KSA. asamar@taibahu.edu.sa
- RT Tan is with National University of Singapore and Yale-NUS College, Singapore. robbyt.tan@nus.edu.sg.
algorithmic approach of [6] was introduced to perform a graph transduction, through the propagation of few labels, by minimising the Laplacian graph energy (2) for the specific case when \( p = 2 \). Subsequent machine learning studies showed that using non-smooth energies with the \( p = 1 \) norm, related to non-local total variation, can achieve better clustering performances [22], but original algorithms only approximated \( p \to 1 \). More advanced optimisation tools were therefore proposed to consider the exact \( p = 1 \) norm for binary [23] or multi-class [24] graph transduction. The normalisation of the operator is nevertheless crucial, as underlined in [25], to ensure within-cluster similarity when the degrees \( d_i \) of the nodes are broadly distributed in the graph.

These motivations drive our approach using the normalised Dirichlet energy (2) based on the graph Laplacian.

**Contributions.** Our work is motivated by the problems of network calibration and the confirmation bias in pseudo-labelling [26], [27], [28], [29], where one seeks that the probability of the predicted label reflects the ground truth correctness likelihood. In particular, in the context of deep learning a current major family of techniques is pseudo-labelling. In this perspective, a current challenge is how to improve poorly calibrated networks for better pseudo-labels [28], [30]. In this work, we argue that energy models can be a powerful alternative for inferring more meaningful pseudo-labels than the ones directly generated from a deep network. With this motivation in mind, we propose a hybrid framework for semi-supervised classification called CREPE network. With this motivation in mind, we propose a hybrid framework for semi-supervised classification called CREPE network (1-Laplacian gRaph Energy for Pseudo-labEls). The core of our proposal is a novel 1-Laplacian graph energy for inferring more certain pseudo-labels, these are then intertwined in an alternating optimisation scheme with a deep network for updating the graph. Our contributions are:

- We propose a hybrid framework for semi-supervised classification, in which we highlight:
  - A new energy model based on the normalised and non-smooth Dirichlet energy (2) based on the graph Laplacian, where we consider the exact \( p = 1 \) norm (energy functional (13) following our scheme (14)). Our functional is based on a carefully chosen class priors to enforce a sufficiently smooth solution, and to strengthen the intrinsic relation between the labelled and unlabelled data.
  - We provide a convergence analysis of our model, and show that the solution trajectory does indeed converge to a non-constant steady point (Proposition 2, Proposition 3). Moreover, we provide a simple yet effective coupling constraint between labels for multi-class problems (Section 4.2).
  - We apply our results (Proposition 4) and derive the connection of our energy model with the principles of pseudo-labeling. We then show how our graph based pseudo-labels can be iteratively updated, in an alternating optimisation scheme, with a deep network.

- We extensively evaluate our model, with numerical and visual results, using six benchmarking datasets from medical and natural images: CIFAR-10/100, Chest-XRay14, CBIS-DDSM, Fashion-MNIST and MiniImageNet. We demonstrate that our technique is able to generalise well to these diverse set-ups, and provides readily competing results against supervised techniques and state-of-the-art results for semi-supervised classification.

### 2 RELATED WORK

The problem of classifying images with scarce annotations has been extensively investigated in the machine learning community. In the literature, semi-supervised learning (SSL) can be broadly divided into three categories: energy models (classic techniques) e.g. [5], [6], [10], [11], deep-learning based methods e.g. [16], [17], [18], [31], and hybrid techniques e.g. [19], [20], [21]. An illustration of each category is displayed in Fig. 1. These categories can be split by different perspectives including graph-based techniques, generative models, pseudo-labelling and consistency regularisation. In this section, we review the existing techniques.

**Energy Models for Semi-Supervised Classification.** Semi-supervised classification has been extensively investigated in the literature, where the underpinning theory of this paradigm has been actively developed since early works e.g. [3], [32], [33], [34], [35]. The solid foundations have been strongly driven by the practical interest of relying less on labelled data in real-world applications such as text analysis [35], [36]. A first family of techniques developed in the area is the energy models [3], [5], [6], where the main idea is to minimise a given energy (a maximum in probability) to infer the labels from the huge amount of unlabelled samples solely relying on deep networks. c) Hybrid techniques use principles from both energy models and deep networks.
large subfamily of techniques is graph based approaches which is the focus of our interest.

Several techniques have been reported following the graph perspective including random walks e.g. [43], [44], harmonic based energy e.g. [5], graph mincut e.g. [9], [45], [46], and spectral techniques e.g. [47], [48]. In most recent works, the authors of [49] used a sparse variant of label propagation under the condition that initial labels are in the proximity of the cluster boundaries. A weighted nonlocal Laplacian energy was introduced in [50], where the authors enforce preservation of the symmetry of the Laplace operator. A kernel clustering approach was used in [51], [52] as an approach for Laplacian regularisation. The Poisson equation on a graph was used in [53] for low label rates classification.

**Deep Semi-Supervised Techniques.** The power of deep learning has been recently applied for semi-supervised classification, which leads the current state-of-the-art performance. A visualisation of this family of techniques is displayed in Fig. 1-b). There exist two major families of techniques in modern semi-supervised techniques: consistency regularisation (aka perturbation-based methods) e.g. [16], [31], [54] and pseudo-labelling e.g. [29], [55], [56]. Consistency regularisation techniques work under the assumption that the model’s performance (output $g(X_u)$, where $X_u$ is the unlabelled data) should not change under any induced $\tau$-perturbation, that is: $g(X_u) = g(X_u + \tau)$. Following this principle, several techniques have been proposed including the works of [16], [17], [18], [56], [57], [58], [59]. A major challenge on these approaches is how to set $\tau$. Different strategies for $\tau$ have been considered in the literature, including mixup augmentations e.g. [60], generative augmentations e.g. [18] and SOTA augmenters e.g. [61], [62]. The core of the performance, of this family of technique, is the use of costly optimisation schemes (e.g. more than 1M training iterations) along with strong augmentations.

The second large family of deep semi-supervised techniques is pseudo-labelling introduced by Lee in [55]. The idea of pseudo-labelling is to generate proxy labels to guide the learning process. Different techniques have been proposed to improve the performance of pseudo-labelling. The use of mix-max feature regularisation was presented in [63]. The authors of [64] proposed a density aware mechanism for improving feature learning and pseudo-label generation. Label propagation using the graph Laplacian with the $p = 2$ case have been proposed in [19] and in combination with clustering regularisation in [20]. Mixup has been shown to offer good performance along with small labels per minibatch [29], and together with graph based pseudo-labels [21]. Certainty mechanisms have also been proposed to improve pseudo-labelling [30], [56].

**Hybrid Techniques and Comparison to Our Work.** Whilst existing techniques either are energy models or deep learning techniques, works simultaneously using these principles, called hybrid techniques, are very recent and scarce (see Fig. 1-c). The existing works are under the family of pseudo-labelling techniques, where energy models have been used for improving performance. The work of [19] adapted the energy model of [6] to an inductive framework using modern deep networks. The same energy model was used in [20] along with clustering regularisation. Most recently, the work of [21] showed high performance by using the energy model of [6] along with a multi-sampling augmentation strategy.

The existing hybrid models have a commonality that is the use of the energy model of [6], where one seeks to minimise (2) for the specific case $p = 2$. The focus of existing works is not the energy model part but rather the development of mechanisms for improving the network performance. In contrast to those works, our approach centers in developing better energy functionals and their theoretical properties to help the learning process. Moreover, existing works have not investigated how more robust energy models impact the performance.

### 3 Preliminaries

This work addresses the problem of semi-supervised classification. In particular, we follow the graph based perspective in semi-supervised learning. Formally, we aim at solving the following problem.

**Problem Statement.** Given a set of samples $X = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ where $x_i \in \mathcal{X}$, we assume that a tiny subset is labelled $D_L = \{(x_i, y_i)\}_{i=1}^L$ with provided labels $\{y_i\}_{i=1}^L \subseteq \mathcal{L} = \{1, ..., L\}$ for $L$ classes, and a large subset is unlabelled $X_u = \{x_i\}_{i=l+1}^n$. We then seek to infer a function $f: \mathcal{X} \mapsto \mathcal{L}$ such that $f$ gets a good estimate for $\{x_i\}_{i=l+1}^n$ with minimum generalisation error.

We address this problem from the hybrid perspective (see Fig. 1-c), where one seeks to combine principles from energy models and deep networks. In particular, our work focus on a hybrid technique from the pseudo-labelling perspective. In deep semi-supervised classification for pseudo-labelling, the main goal is to solve a loss that relates the labelled and unlabelled sets, whose general form reads:

$$\min_{\theta} \sum_{(x,y) \in D_L} L_S(x, y; \theta) + \gamma \sum_{x \in X_u} L_U(x, \hat{y}; \theta),$$

where the two terms $L_S$ and $L_U$ handle the labelled and unlabelled set respectively, $\hat{y}$ is the network parameters, $y$ are pseudo-labels, and $\gamma$ is a positive parameter weighting the importance of each term. In the body of literature, the main difference between existing works is the way to define $L_U$.
for example using a pseudo-labelling strategy or consistency regularisation.

A current major challenge is how to better handle issues relating to network calibration and confirmation bias e.g. [26], [27], [28], [29]. In the context of pseudo-labelling, hybrid techniques e.g. [19], [20], [21] have shown that one can mitigate such issues by inferring pseudo-labels from an energy model and then combine them with deep networks rather than predict them directly from a deep network. However, existing hybrid techniques have only focused on designing mechanism for the networks and the investigation of better energy models are to be investigated.

With previous motivation in mind, we seek to design better functionals for improving the inference of pseudo-labels. To do this, we follow a graph based perspective for inferring more certain pseudo-labels. In this work, we consider functions \( u \in \mathbb{R}^n \), defined over a set \( \mathcal{N} \) of \( n \) nodes. Our main points of interest are convex and absolutely \( p \)-homogeneous (i.e. \( J(\alpha u) = |\alpha|^p J(u) \)) non-local functionals, defined on \( u \), of the particular form:

\[
D_p(u) = \sum_{ij} w_{ij} \left\| \frac{u_i}{d_i^{1/p}} - \frac{u_j}{d_j^{1/p}} \right\|^p,
\]

with weights \( w_{ij} = w_{ji} \geq 0 \) taken such that the vector \( d \in \mathbb{R}^n \) has non-null entries satisfying: \( d_i = \sum_j w_{ij} > 0 \). This energy acts on the graph defined by nodes \( \mathcal{N} \) and weights \( w_{ij} \). With respect to classical Dirichlet energies associated to the graph \( p \)-Laplacian [23], [24], [65], [66], it includes a normalisation through rescaling with the degree of the node.

In this work, we focus our attention to the non smooth case of better energy models are to be investigated.

Proposition 1. For all \( \phi \in \partial J \), with \( J \) defined in (3), one has

\[
\langle \phi, d \rangle = 0.
\]

Proof. Observing that \( d = D1 \) and using (5) we have that

\[
\langle \phi, d \rangle = \langle D^{-1}W^T z, D1_n \rangle.
\]

Since the weights satisfy \( w_{ij} = w_{ji} \), then for all \( z \in \mathbb{R}^m \):

\[
W^T z, 1_n \rangle = \sum_i \sum_j w_{ij} (z_i - z_j) = \sum_i \sum_{j > i} w_{ij} (z_i - z_j - z_i + z_j) = 0.
\]

\[\square\]

Eigenfunctions. Eigenfunctions of any convex functional \( J \) satisfy \( \lambda u \in \partial J(u) \). For \( J \) being the nonlocal total variation, (i.e. when \( d_i \) is constant), eigenfunctions are known to be essential tools to provide a relevant clustering of the graph [25]. Methods [22], [67], [68], [69], [70] have thus been designed to estimate such eigenfunctions through the local minimisation of the Rayleigh quotient, which reads:

\[
\min_{|u|_2 = 1} \frac{J(u)}{H(u)},
\]

with another absolutely one homogeneous function \( H \), that is typically a norm. Taking \( H(u) = |u|_2 \) as the \( \ell_2 \) norm, one can recover eigenfunctions of \( J \). For \( H(u) = |u|_1 \) being the \( \ell_1 \) norm, one can also compute bi-valued functions \( u \) that are local minima of (6) and eigenfunctions of \( J \) [71]. Being bivalued, these estimations can easily be used to realise a partition of the domain. These schemes also relate to the Cheeger cut of the graph induced by nodes \( u_i \) and edges \( w_{ij} \). Balanced cuts can also be obtained by considering \( \hat{H}(u) = \frac{1}{|u|} \).

A last point to underline comes from Proposition 1 that states that eigenfunctions \( \lambda u \in \partial J(u) \) should be orthogonal to \( d \). It is thus important to design schemes that ensure this property.
4 Improving Pseudo-Labelling with the 1-Laplacian Graph Energy

This section describes our proposed energy model that fits into a hybrid framework that we called CREPE Model (1-Laplacian gRaph Energy for Pseudo-labEls). As illustrated in Fig. 3, there are two main components in hybrid techniques: a deep network and an energy model. In this work and unlike existing hybrid techniques that focused on designing better mechanisms for improving the network performance, we focus on developing better energy functionals and their theoretical properties (see green box from Fig. 3). This section describes three key parts: i) the convergence of our energy model, ii) the definition of our coupling constrain for the multi-class problem and iii) our multi-class energy flow for pseudo-labelling.

Core Idea. We seek to infer better pseudo-labels using an energy model (outside a deep network) instead of generating them directly from a deep network. To do this, we introduce a 1-Laplacian graph energy (see green box from Fig. 3), which is detailed next.

4.1 Convergence Analysis

In the following, we will denote the value of function $u$ at node $x$ by $u(x)$ and the value of $u$ at iteration $k$ as $u_k$. In order to realise a binary partition of the domain of the graph $\mathcal{N}$ through the minimisation of the quotient $R(u) = J(u)/H(u)$, we adapt the method of [71] to incorporate the scaling $d(x)$ of (3) and consider the semi-explicit scheme:

$$
\begin{align*}
\left\{ \begin{array}{l}
\frac{u_{k+1/2} - u_k}{\delta t} = \frac{J(u_k)}{H(u_k)} (q_k - \hat{q}_k) - \phi_{k+1/2}, \\
\frac{u_k}{\|u_k\|_2} = \frac{\phi_k}{\|\phi_k\|_2}
\end{array} \right.
\end{align*}
$$

(7)

with $\phi_{k+1/2} \in \partial J(u_{k+1/2})$, $q \in \partial H(u_k)$, $\hat{q}_k = \langle d, q \rangle / \| q \|_2$. We recall that both $J$ and $H$ are absolutely one homogeneous and satisfy (4). Since $\langle \phi, d \rangle = 0$, $\forall \phi \in \partial J$, the shift with $\hat{q}_k$ is necessary to show the convergence of the scheme (7) as we have $u_k \rightarrow u^* = \frac{\partial J(u^*)}{\partial u^*} (q^* - \hat{q}^*) = \phi^*$, for $\phi^* \in \partial J(u^*)$ and $q^* \in \partial H(u^*)$.

Such sequence $u_k$ satisfies the following properties.

**Proposition 2.** For $(u_0, d) = 0$, the trajectory $u_k$ given by (7) satisfies:

1. $(u_{k+1}, d) = 0$,
2. $\|u_{k+1/2}\|_2 \geq \|u_k\|_2$,
3. $R(u_k)$ is non-increasing,
4. $H(u_{k+1/2}) \leq \kappa < +\infty$.

Proof. In this proof, we use the fact that $u_{k+1/2}$ defined in (7) is the unique minimiser of:

$$
F_k(u) = \frac{1}{2\delta t}|u - u_k|_2^2 + R(u_k)\langle q_k - \hat{q}_k, u \rangle + J(u).
$$

(8)

For $(u_k, d) = 0$, we have

$$
\langle u_{k+1/2}, d \rangle = \langle u_k, d \rangle + \delta t (R(u_k)\langle (q_k - \hat{q}_k), d \rangle - \langle \phi_{k+1/2}, d \rangle)
$$

$$
= \delta t R(u_k) \langle (q_k - \hat{q}_k), d \rangle
$$

$$
= 0,
$$

where we used Proposition 1 in the right part of the previous relation to get $\langle \phi_{k+1/2}, d \rangle = 0$. We conclude with the fact that $u_{k+1}$ is a rescaling of $u_{k+1/2}$.

Since $H$ is a norm, it is absolutely one homogeneous and $q_k \in \partial H(u_k) \Rightarrow H(u_k) = \langle q_k, u_k \rangle$. Next, we observe that $J(u_k) = \sup_{\delta \in \partial J} \langle \delta, u_k \rangle \geq \langle \phi_{k+1/2}, u_k \rangle$ and we get

$$
\langle u_{k+1/2}, u_k \rangle = \|u_k\|_2^2 + \delta t (R(u_k)\langle q_k - \hat{q}_k, u \rangle - \langle \phi_{k+1/2}, u_k \rangle)
$$

$$
\geq \|u_k\|_2^2 + \delta t (H(u_k) - R(u_k)\langle \hat{q}_k, u \rangle - J(u_k))
$$

$$
\geq \|u_k\|_2^2 - \delta t R(u_k) \langle d, q_k \rangle \langle d, d \rangle
$$

$$
\geq \|u_k\|_2^2.
$$

We then conclude with the fact that $\langle u_{k+1/2}, u_k \rangle \leq \|u_{k+1/2}\|_2 \|u_k\|_2$.

Since $(u_k, d) = 0$ for all $k$ and $\hat{q}_k = \langle d, q_k \rangle / \| q \|_2 d$, then $(\hat{q}_k, u_{k+1/2}) = (\hat{q}, u_k) = 0$. Next, we recall that $H(u_{k+1/2}) = \sup_{q \in \partial H} \langle q, u_{k+1/2} \rangle \geq \langle \hat{q}, u_{k+1/2} \rangle$. Hence we have

$$
F_k(u_{k+1/2}) \leq F(u_k)
$$

$$
\frac{1}{2\delta t}\|u_{k+1/2} - u_k\|_2^2 - R(u_k)\langle q_k, u_{k+1/2} \rangle + J(u_{k+1/2}) \leq 0
$$

(9)

$$
\frac{1}{2\delta t}\|u_{k+1/2} - u_k\|_2^2 + J(u_{k+1/2}) \leq R(u_k)H(u_{k+1/2})
$$

$$
R(u_{k+1/2}) \leq R(u_k)
$$

$$
R(u_{k+1}) \leq R(u_k)
$$

where the final rescaling with $\|u_{k+1/2}\|_2$ is possible since $J$ and $H$ are absolutely one homogeneous functions.

In the finite dimension setting, there exists $K_J, K_H < \infty$ such that $\|p\| \leq K_J$ and $\|q\| \leq K_H$ for an absolutely one homogeneous functionals $J$ defined in (1) and a norm $H$. Then one has

$$
\|u_{k+1/2}\|_2 = u_k + \delta t \left( \frac{J(u_k)}{H(u_k)} (q_k - \hat{q}_k) - \phi_{k+1/2} \right)
$$

$$
\|u_{k+1/2}\|_2^2 \leq \langle u_{k+1/2}, d \rangle + \delta t \left( \frac{J(u_k)}{H(u_k)} \langle q_k, u_{k+1/2} \rangle - \langle \phi_{k+1/2} \rangle \right)
$$

$$
\|u_{k+1/2}\|_2 \leq \|u_k\|_2 + \delta t \left( \frac{J(u_k)}{H(u_k)} K_H + K_J \right)
$$

From the equivalence of norms in finite dimensions, there exists $0 < \kappa < \infty$ such that $H(u_{k+1/2}) \leq \kappa$.

Hence, we can show the convergence of the trajectory.

**Proposition 3.** The sequence $u_k$ defined in (7) converges to a non-constant steady point $u^*$.

Proof. As $u_{k+1/2}$ is the unique minimiser of $F_k$ in (8), as $F_k(u_k) = 0$, and as $\langle q_k - \hat{q}_k, u_{k+1/2} \rangle \leq H(u_{k+1/2})$, we get:

$$
\frac{1}{2\delta t H(u_{k+1/2})}\|u_{k+1/2} - u_k\|_2^2 + R(u_{k+1}) \leq R(u_k).
$$

(10)

Since $u_{k+1}$ is the orthogonal projection of $u_{k+1/2}$ on the $L_2$ ball then $\|u_{k+1} - u_k\|_2 \leq \|u_{k+1/2} - u_k\|_2$. Finally, from statement 4 of Proposition 2, we have that $1/H(u_{k+1/2}) \geq 1/\kappa$. We then sum relation (10) from 0 to $K$ and deduce that:
Contrary to the simplex constraint, it is compatible
Projection on this linear constraint is explicit with a sim-
pling instead of classical simplex [24], [73], [74] or orthogo-
u for pseudo-labelling.

In previous section, we provide the convergence analysis and
coupling constraint of our energy model. In this section, we
between the coupled functions
satisfies
where \( \hat{q}_k \in \partial H(u_k^L) \) and \( \chi_C \) is the characteristic function of
the constraints (11). Starting from an initial point \( u_0 \)
that satisfies the constraint \( (\chi_C(u_0) = 0) \) and has been
normalised \( |u_0|^2 = \sum_{l=1}^{L} |u_0|^2 = 1 \) , the scheme we
consider reads:

\[
\begin{align*}
\{ u_{k+1/2} &= \arg\min_u F_k^L(u) \\
 u_{k+1} &= \frac{u_{k+1/2}}{|u_{k+1/2}|^2}
\end{align*}
\]  

In practice, if for some \( l \), \( u_{k+1/2}^l \) vanishes, then we define
\( R(u_{k+1}) = 0 \) for the next iteration. With such assumptions,
the sequence \( u_k \) have the following properties.

**Proposition 4.** For \( (u_k^l, d) = 0, l = 1 \cdots L \), the trajectory \( u_k \)
given by (14) satisfies

1. \( (u_k^l, d) = 0 \),
2. \( |u_k|^2 \leq |u_{k+1/2}|^2 \leq \kappa < \infty \),
3. \( \sum_{l=1}^{L} H(u_{k+1}) (R(u_{k+1}) - R(u_k^l)) \leq -\frac{1}{2\delta t\kappa} |u_{k+1} - u_k|^2 \).

**Proof.** The scheme reads

\[
\begin{align*}
 u_{k+1/2} &= \arg\min_u F_k^L(u) := \frac{1}{2\delta t} |u - u_k|^2 + \sum_{l=1}^{L} J(u^l) \\
 &- \sum_{l=1}^{L} R(u_k^l)(q_k^l - \hat{q}_k^l, u^l) + \chi_C(u) \\
 u_{k+1} &= \frac{u_{k+1/2}}{|u_{k+1/2}|^2}
\end{align*}
\]

The Karush–Kuhn–Tucker conditions of the above problem
states that there exist \( \phi_{k+1/2}^l \) and \( r_{k+1/2} \) such that

\[
u_{k+1/2}^l = u_k^l + \delta t \left( R(u_k^l)(q_k^l - \hat{q}_k^l) - \phi_{k+1/2}^l - r_{k+1/2} \right)
\]

where \( \phi_{k+1/2}^l \in \partial J(u_{k+1/2}^l) \) and \( r_{k+1/2} \) is a Lagrange
multiplier independent of \( l \) for the linear constraint \( \chi_C \). The
point \( u_{k+1/2} \) in the above scheme corresponds to the global
minimiser of \( F_k^L(u) \).

For \( (u_k^l, d) = 0 \), and following point 1 of Proposition 2, we have

\[
\begin{align*}
\langle u_{k+1/2}^l, d \rangle &= \langle u_k^l, d \rangle + \delta t (R(u_k^l)(q_k^l - \hat{q}_k^l), d) \\
&- \langle \phi_{k+1/2}^l, d \rangle - \langle r_{k+1/2}, d \rangle \\
&= -\langle r_{k+1/2}, d \rangle.
\end{align*}
\]
Next, as $u_{k+1/2} \in C$, we have $\sum_l u_{k+1/2}^l(x) = 0, \forall x \in \mathcal{N}$ and obtain:

\[
\sum_{l=1}^L \langle u_{k+1/2}^l, d \rangle = -\sum_{l=1}^L \langle r_{k+1/2}^l, d \rangle
\]

\[
\sum_{l=1}^L \sum_{x \in \mathcal{N}} u_{k+1/2}^l(x) d(x) = -L \langle r_{k+1/2}, d \rangle
\]

\[
\sum_{x \in \mathcal{N}} d(x) \left( \sum_{l=1}^L u_{k+1/2}^l(x) \right) = -L \langle r_{k+1/2}, d \rangle
\]

0 = \langle r_{k+1/2}, d \rangle.

We have

\[
\langle u_{k+1/2}^l, u_{k}^l \rangle = \|u_{k}^l\|^2 + \delta \left( \frac{R(u_{k}^l)}{\|u_{k}^l\|^2} - \frac{\delta}{\|u_{k+1/2}^l\|^2} \right).
\]

We follow the point 2 of Proposition 2 to first get:

\[
\langle u_{k+1/2}^l, u_{k+1/2}^l \rangle \geq \|u_{k+1/2}^l\|^2 - \langle r_{k+1/2}, u_{k+1/2}^l \rangle
\]

\[
+ R(u_{k+1/2}^l) q_k - \delta q_k, u_{k+1/2}^l
\]

Summing on $l$, we get

\[
\|u_{k+1/2}\|^2 \leq \|u_k\|^2 + \delta \left( \sum_{l=1}^L R(u_k^l) q_k - \delta \sum_{l=1}^L q_k, u_{k+1/2}^l \right).
\]

Notice that we defined $R(u_k^l) = 0$ for $u_k^l = 0$. As $H$ is a norm, the equivalence of norm in finite dimensions implies that $\|u_k^l\|^2 H(u_k^l)$ is bounded by some constant $c < \infty$. We then have $\|u_{k+1/2}\|^2 \leq \frac{1}{1+\delta} K(1+LKH)$.

Since $u_{k+1/2}$ is the global minimiser of (10), then:

\[
F^{L}_k(u_{k+1/2}) \leq F^{L}_k(u_k)
\]

\[
\frac{1}{2\delta} \|u_{k+1/2} - u_k\|^2 + \sum_{l=1}^L J(u_k^l) \leq \sum_{l=1}^L R(u_k^l) q_k - \delta q_k, u_{k+1/2}^l
\]

\[
\frac{1}{2\delta} \|u_{k+1/2} - u_k\|^2 + \sum_{l=1}^L J(u_k^l) \leq \sum_{l=1}^L R(u_k^l) H(u_{k+1/2}^l)
\]

\[
\|u_{k+1/2}\| \leq \sum_{l=1}^L J(u_k^l) \leq \sum_{l=1}^L R(u_k^l) H(u_{k+1/2}^l)
\]

\[
\|u_{k+1/2}\| \leq \sum_{l=1}^L H(u_k^l) \left( R(u_k^l) - R(u_k^l) \right) \leq \frac{1}{2\delta} \|u_{k+1/2} - u_k\|^2
\]

Point 3 of Proposition 4 contains weights $H(u_{k+1}^l)$ that prevent from showing the exact decrease of the sum of ratios. This is in line with the results in [24]. To ensure the decrease of the sum of ratios $\sum_{l=1}^L J(u_k^l)/H(u_k^l)$, it is possible to introduce auxiliary variables to deal with individual ratio decrease, as in [73]. The involved sub-problem at each iteration $k$ is nevertheless more complex to solve.

### 4.4 Introducing Label Priors

The partitioning process induced by the scheme (14) so far does not integrate any label information. As we are working in a semi-supervised setting, we consider given a small subsets of labelled nodes $\mathcal{N}^l \subset \mathcal{N}$ (with $|\mathcal{N}^l| << |\mathcal{N}|$) belonging to each cluster $i$, with $\mathcal{N}^l \cap \mathcal{N}^j = \emptyset$. Denoting as $\mathcal{L} = \cup_{i \in \mathcal{N}^l} \mathcal{N}^l$, the objective is to propagate the prior information in the graph in order to infer pseudo-labels for the remaining nodes $x \in N \setminus \mathcal{L}$. To that end, we simply have to modify the coupling constraint $C$ in (11) as

\[
C : \begin{cases} u_k, & \text{s.t.} \sum_{l=1}^L u_k^l(x) = 0 \quad \text{if } x \in \mathcal{N} \setminus \mathcal{L} \\text{and } u_k^l(x) \geq \epsilon \quad \text{if } x \in \mathcal{N}^l \\text{and } u_k^l(x) \leq -\epsilon, \forall l', \exists l \quad \text{if } x \in \mathcal{L} \setminus \mathcal{N}^l \end{cases}.
\]

With such constraint, clusters can no longer vanish or merge since they all contain different active nodes $x \in N$ satisfying $u_k(x) > 0$. The same scheme (14) with the new constraint set (15) can be applied to propagate these labels. Once it has converged, the inferred pseudo-label of each unlabelled node $x \in N \setminus \mathcal{L}$ is taken as:

\[
L(x) = \arg \max_{i \in \{1, \ldots, L\}} u_k^i(x). \quad (16)
\]

Soft labelling can either be obtained by considering all the clusters with non-negative weights $\mathcal{I}(x) = \{l, u_k^l(x) \geq 0\} \neq \emptyset$ and with relative weights $u_k^l(x) = u_k(x)/\sum_{l \in \mathcal{I}(x)} u_k^l(x)$, with the convention that $u_k(x) = 1/L$, in the case that $u_k(x) = 0$ for all $l = 1 \cdots L$ (which has never been observed in our experiments). For notation purposes and following the notation in (1), we denote the output of (16) as $\hat{y}_i$, having $\hat{Y} = \{\hat{y}_i\}_{i=1}^L$. The parameter $\epsilon$ in (15) is set to a small numerical value.

### 4.5 Hybrid Framework

We now fit our energy model into a hybrid framework as displayed in Fig. 3. We perform an alternating optimisation between (1) and (14). The process (14), that provides pseudo-labels $\hat{y}$, is extensively described in previous subsections (as our main contribution). The functional (1) is now detailed. For the first term in (1), we use a cross entropy loss with a weighting parameter as imbalance class strategy. We follow a standard strategy e.g. [76], [77] such that the parameter is inversely proportional to the number of samples for class $k : \beta_k \propto 1/|E_n|$ where $E_n$ is the total number of samples. The second term in (1) involves the inferred pseudo-labels updates $\hat{y}$ in a cross entropy loss, along with a dual weighting parameter $\beta_{\varphi}$ where $\varphi$ is a measure of the uncertainty referring to the entropy. The remaining of the experiments follows this alternating optimisation. The choice for $f_0$ is discussed in the experimental results section.
5 Experimental Results

This section focuses on the detailed description of the experiments that we conducted to evaluate our proposed approach.

5.1 Data Description

We extensively evaluate our approach using six very diverse datasets. Firstly, we use the Fashion-MNIST [78] dataset. The dataset is composed of 70k grayscale images containing 10 classes from fashion items. To further support the generalisation and robustness of our technique, we use two major complex datasets from the medical domain. The ChestX-ray14 dataset [79] is composed of 112,120 frontal chest view X-ray with size of $1024 \times 1024$ and 14 classes reflecting diverse pathologies. The CBIS-DDSM dataset [80], composed of 3,103 mammography images with a mean size of $3138 \times 5220$, contains normal, benign, and malignant cases with verified pathology information. Finally, we use three natural image datasets. The CIFAR-10 and CIFAR-100 dataset [81] that contains 60k colour images of size $32 \times 32$ with 10 and 100 different classes respectively. Finally, the Mini-ImageNet [82] dataset consisting of 60k colour images with of size $64 \times 64$ and 100 classes.

5.2 Evaluation Protocol

We design the following evaluation scheme to validate our theory.

Baseline Comparison against existing energy models. As the core contribution of this work is a new graph based energy model, we first compared our technique against existing graph-based energy methods: Harmonic Gaussian (HG) [5], Local to Global Consistency (LGC) [6], Lazy Random Walks (LRW) [44], Sparse Label Propagation (SLP) [49], Weighted Nonlocal Laplacian (WNLL) [50], Centered Kernel (CK) [51], [52] and Poisson Learning (PoL) [53]. To solely evaluate the impact of these energy models, we used the same network architecture and substituted each method as the energy model, (i.e. only green box of Fig. 3). We run the experiments for all techniques under the same conditions by constructing a k-NN graph with $k = 20$ using the features extracted from a 13-Layer Network. We remark that these set of experiments are to purely compare energy models and not networks Fig. 3. We use different label % counts [1,2,5,10,20,30], and report the mean error and standard deviation over randomly select the labelled samples over twenty repeated times (20 different splits).

Comparison against SOTA Techniques. For our full model we compare to the state-of-the-art for each of the differing domains

Medical Datasets. For the ChestX-ray14 dataset [79], we firstly compared against the SOTA supervised techniques of [79], [83], [84], [85], [86], [87], [88], [89], [90] using the official partition of the dataset (70% labelled data) against ours using 20% of labelled data. Moreover, we compared against the SOTA semi-supervised techniques of [17], [91], [92], [93]. All semi-supervised techniques are reported using 20% of labels. We also provide comparison with existing techniques [80] [94], [95], [96], [97] on the CBIS-DDSM dataset. The quality check is performed following standard convention in the medical domain by a ROC analysis using the area under the curve (AUC).

Natural Image Datasets. Finally, we report results against the SOTA semi-supervised techniques for natural image datasets: II−Model and Temporal Ensembling [16], Mean Teacher (MT) [17], VAT [18], II−SNTG [98], MT+fast-SWA [57], MT+ICT [58], Dual Student [59], MUSCLE+MT+LP [54], MT+TSSDL [63], MT+LP [19], CycleCluster [20], DAG [64], UPS [30], PL-Mixup [29], LaplaceNet [21], UDA [99], SimPLE [56], FixMatch [31]. We evaluate the quality of the classifiers by reporting the error rate and standard deviation over five runs and for a range number of labelled samples.

5.3 Implementation Details

We set the architecture for $f_\theta$ (i.e. red box of Fig. 3) as follows. For the medical datasets we use a ResNet-18 [2]. For the natural image datasets, we ran experiments with three different networks. For CIFAR-10 and CIFAR-100, we divided our experiments into two parts. For the first part, we use a 13-Layer Network for a fair comparison as existing approaches run under this architecture. For the second part and motivated by the work of [100], we compare against the most recent techniques under exactly the same conditions which includes the optimiser, RandAugment implementation and a WideResNet-28-2 (WRN-28-2). Finally, we use a ResNet-18 for the Mini-ImageNet dataset as fair comparison for existing techniques. For the graph generation, a k-NN graph with $k = 20$ is constructed using the features from each respective architecture - Fig. 4 displays examples of generated graphs for two selected datasets. For our approach, we set the number of epochs of 310 and a weight decay of $2 \times 10^{-4}$. The learning rate is set to 5e-2 and with a scheduled cosine annealing. We use as optimiser stochastic gradient descent (SGD) and implement our code in PyTorch.

5.4 Results & Discussion.

In this section, we report and discuss the results and comparison of our proposed technique.

How Good is our Energy Model? We start by evaluating the performance of our energy model. To do this, we ran a set of comparisons of our technique against existing energy models including recent ones. For a fair comparison all techniques were fed with the same graph (constructed as detailed in previous subsection). The results are displayed in
### Table 1
Numerical comparison of our approach vs other energy-based approaches. The values are computed from 20 runs as the average of error rate ± standard deviation over several label counts. The best results are highlighted in green whilst the second best ones are highlighted in red.

| Technique | 1% | 2% | 5% | 10% | 20% | 30% |
|-----------|----|----|----|-----|-----|-----|
| Harmonic Gaussian (HG) [5] | 18.97±0.34 | 17.96±0.23 | 16.35±0.10 | 15.33±0.10 | 14.48±0.11 | 14.09±0.08 |
| Local to Global Consistency (LGC) [6] | 18.65±0.59 | 17.81±0.31 | 16.41±0.12 | 15.20±0.15 | 14.33±0.13 | 13.68±0.06 |
| Lazy Random Walks (LRW) [44] | 19.09±0.25 | 17.84±0.17 | 16.38±0.06 | 15.69±0.10 | 15.35±0.10 | 15.19±0.07 |
| Sparse Label Propagation (SLP) [49] | 78.18±1.37 | 50.16±3.81 | 44.18±0.37 | 25.17±1.75 | 19.03±1.75 | 13.16±0.11 |
| Weighted Nonlocal Laplacian (WNL) [50] | 19.68±0.17 | 18.96±0.28 | 17.41±0.12 | 16.15±0.17 | 14.93±0.13 | 14.32±0.09 |
| Centered Kernel (CK) [51], [52] | 31.16±1.28 | 24.29±0.61 | 20.30±0.11 | 18.62±0.24 | 16.54±0.12 | 15.38±0.07 |
| Poisson Learning (PoL) [53] | 20.60±1.37 | 19.89±3.81 | 19.17±0.37 | 19.03±1.75 | 18.82±0.71 | 18.87±0.41 |
| Ours | 17.62±0.21 | 16.05±0.24 | 14.21±0.08 | 13.16±0.11 | 12.31±0.09 | 11.82±0.04 |

### Table 2
Performance comparison of our approach (20% of labelled data) against SOTA fully-supervised techniques (70% of labelled data). We reported the AUC per class and average AUC over all 14 pathologies. All compared techniques are performed using the official data split. The results in bold denote the highest performance.

#### Table 3
Numerical comparison of our technique and existing semi-supervised approaches for the ChestX-ray14 dataset. All techniques use 20% of labelled data, and the reported results reflect the AUC per class and average. The best results are highlighted in green.

#### Table 4
AUC performance comparison of existing SOTA supervised techniques and our technique for the CBIS-DDSM dataset. The best results are marked in green whilst the second best in red.

Table 1, which reports the error rate averaged over 20 runs and the standard deviation under different % of labelled data. In a closer look at the results, we can observe that our approach reports the lowest error rate for all label counts whilst LGC [6] ranked second. The techniques of CK [51] and SLP [49] failed to be robust in the low label regime and needed a higher number of labels to improve the performance than the compared techniques. A similar performance behaviour was observed in the techniques of HG [5], LGC [6], LRW [44] and WNL [50]. In contrast to the compared techniques, the performance of PoL [53] was not improved when more labels are considered. Our
We now evaluate our full hybrid framework (see Fig. 3). We start by using the ChestX-ray14 benchmarking dataset [79]. We first compared our approach against the SOTA supervised techniques, they assume a large corpus of annotated data (70%) whilst our technique reports the performance using 20% of labels. The results are reported in Table 2 displaying the AUC per class and the average over all classes. By inspection we can observe that our technique readily competes with existing deep supervised techniques in per class performance. Overall, our technique outperformed almost all existing techniques and places second behind the work of [90]. However, we remark that our technique is using far less labels (only 20%) than all compared techniques (fully supervised 70% of labelled data).

We also compared our proposed technique against existing semi-supervised models for medical imaging. The results are reported in Table 3. All results are produced using 20% of labelled data, and the table displays the AUC per class and average over all classes. We observe that our technique reports the best AUC per class in majority of the pathologies, with a best overall score (see scores highlighted in green). We further support the performance of our method with another challenging medical dataset the CBIS-DDSM dataset [80] for mammograms classification. We compare our approach against existing techniques for such dataset and report the results in Table 4. The compared techniques are deep supervised techniques, and to the best of our knowledge, there exists no modern semi-supervised techniques for this dataset to compare with. For the supervised techniques the official partition is used (i.e. 85% of labelled data). For our technique, we reported the AUC as result of an average of 5 split runs using 35% and 40%

| TECHNIQUE | LABELLED SAMPLES (CIFAR-10) | LABELLED SAMPLES (CIFAR-100) |
|-----------|-----------------------------|-----------------------------|
|           | 1k                          | 2k                          | 4k                          | 10k                        |
| Fully Supervised | 26.60±0.22                  | 19.53±0.12                  | 14.02±0.10                  | 53.10±0.34                 | 36.59±0.47               |

**TABLE 5**

Performance comparison using CIFAR-10/100 reporting error rate ± standard deviation. The consistency regularisation and pseudo-labelling techniques were run under the same code-base using a 13-Layer architecture, and reflect the error rate and standard deviation. The results in bold denotes the best performance.

| TECHNIQUE                  | LABELLED SAMPLES (CIFAR-10) | LABELLED SAMPLES (CIFAR-100) |
|----------------------------|-----------------------------|-----------------------------|
|                             | 2k                          | 4k                          | 10k                        |
| UDA [99]                   | 5.61±0.16                   | 5.40±0.19                   | 36.19±0.39                 | 31.49±0.19                 |
| FixMatch [31]              | 5.42±0.11                   | 5.30±0.08                   | 34.87±0.17                 | 30.89±0.18                 |
| SimPLE [56]                | 5.27±0.18                   | 5.33±0.20                   | 34.75±0.16                 | 29.18±0.25                 |
| LaplaceNet [21]            | 4.71±0.05                   | 4.35±0.10                   | 33.16±0.22                 | 27.49±0.22                 |
| CREPE (Ours)               | 4.33±0.09                   | 4.16±0.11                   | 32.21±0.18                 | 26.14±0.24                 |

**TABLE 6**

Comparison of our technique against existing semi-supervised techniques (consistency regularisation and pseudo-labelling family of techniques) along with the fully supervised baseline for CIFAR-10 and CIFAR-100. All the results are derived from using a 13-Layer architecture, and reflect the error rate and standard deviation. The results in bold denotes the best performance.
of labelled data. From the results, our technique produced readily compared performance whilst using only a fraction of labels, and reported the highest AUC using less than half of the labels than the compared techniques.

**Comparison with SOTA Semi-Supervised Models.** For our final set of experiments, we compared our technique against existing semi-supervised techniques for natural images where numerous methods have been proposed. For a fair comparison, we first provide a comparison using a 13-Layer Network, which is the most widely used network in semi-supervised classification. The results, in terms of error rate, are presented in Table 5. We observe that our technique provides a substantial improvement in performance with respect to consistency regularisation techniques for both CIFAR-10/100 datasets. In terms of existing pseudo-labelling techniques, our technique provides a significant margin of improvement. We namely obtain the lowest error rates for all the label counts and for both datasets.

Most recent and current SOTA techniques are based on more complex optimisation schemes scaling to more modern networks. Therefore, we also provide results against the techniques of: UDA [99], FixMatch [31], SimPLE [56] and LaplaceNet [21]. To do this and following [100] for a fair comparison, we ran those set of techniques under the same code-based (i.e. the same implementation for the augmentations (RandAugment), optimiser and network architecture) using the same backbone a WRN-28-2. The performance comparison in terms of error rate is reported in Table 6 using [2k, 4k] and [4k, 10k] labels for CIFAR-10/100 respectively. Our technique reported the lowest error rate for all label counts and both datasets. We thus observe a significant performance improvement with respect to consistency regularisation techniques for larger class number (CIFAR-100). Finally to further support the generalisation of our technique, we report results for Mini-ImageNet in Table 7. In this experiment all methods use a ResNet-18. We highlight that for this complex dataset, our technique reports a substantial performance improvement ([3%, 37%]).

**Our Energy Model vs Deep Network for Pseudo-labelling.** Our graph energy model offers an alternative to the inherent problem of network calibration and confirmation bias for pseudo-labelling. To further support this argument and our extensive experiments, we provide a set of experiments to showcase the advantages of our energy model vs deep network for pseudo-label generation. We use for CIFAR-10/100 a 13 Layer Network whilst for Mini-ImageNet a ResNet-18. To do this, we run our framework from Section 4.5 with our energy model, and without it and allowing the network, directly from \( f_\theta \), to generate the pseudo-labels. The results are displayed in Fig. 5. In a closer look at the results, we can observe that in Fig. 5-a that integrating our energy model encourage better pseudo-labels, which is reflected in having better performance than the deep network. This behaviour is consistently observed across all compared datasets. We also illustrate the certainty of the pseudo-labels over selected epochs from our approach in Figs. 5-ac. We observe that our model enforces constant control on the level of certainty of the inferred pseudo-labels over the learning process. This effect can be seen in the plots, where the green shaded area, that reflects the correctness of the pseudo-labels with respect to the ground truth, increases with the evolution of the epochs; whilst the number of incorrect pseudo-labels (see red area) decrease.

**A Better Energy Model.** Another key motivation of our work is the need for a robust energy model as existing hybrid techniques [19], [20], [21] have as commonality the

---

**Fig. 5.** (a) Error rate comparison for pseudo-label generation between our energy model vs the deep network for Mini-ImageNet, CIFAR-10/100. (b)-(c) Certainty \( \psi \) of the pseudo-labels correctness (green) along with those incorrect (red) for two selected epochs for CIFAR-100 (4k labels) and Mini-ImageNet (10k labels) correspondingly.

**Fig. 6.** Average AUC comparison of hybrid models performance using our energy model (red line) and that of [6] (blue line). (a) reports the comparison performance for different % of labels for the ChestX-ray14 dataset whilst b) for the CBIS-DDSM dataset.
use of the energy model of that [6]. We firstly showed in Tables 5, 6, 7 that our approach outperforms those existing hybrid techniques. The use of such energy model is motivated by its performance as ranked second in Table 1. To further support our results from that table, we run an additional set of experiments to further evaluate the gain of our energy model vs [6] for more complex data – the ChestX-ray14 (see plot (a)) and CBIS-DDSM (see plot (b)) medical datasets. To do this, we run the hybrid framework from Section 4.5 with our energy model and the energy of [6] for different label rates – that is, changing the green block from Fig. 3. The results are displayed in Fig. 6 in terms of average AUC over the classes. We observe that our technique consistently outperforms that of LGC [6] for all label rates and both datasets. More precisely, we report a performance improvement in the range of 10% to 16% on the different label rates. We also can observe that the both graphical approaches reach a point where more labels are not providing a significant performance improvement. This is an expected behaviour and follows the findings of several early works where the graphical tranductivity bonus is not longer effective as the nature of working on low label rates [34].

Overall Remarks. From our results, we now summarise our main highlights over existing techniques:

flammable energies for better pseudo-labels. From our experiments, we observe that energy models are a strong approach for pseudo-labelling. The intuition behind our technique’s performance is that our energy model allows an explicit control and update of the predictive uncertainty on the pseudo-labels. By contrast, the compared techniques solely rely on the deep network to get the output without any guarantee or clear understanding on the correctness likelihood of the pseudo-labels. 

- Advantages of our Hybrid Model. Unlike existing energy models, our framework takes advantages of both a robust energy model and deep learning principles. In contrast to pure deep learning techniques, our work offers several mathematical properties such as convergence of the scheme and a better understanding of the technique’s behaviour. Finally, in comparison to existing hybrid techniques that use existing energy models and focus on new deep learning mechanism, we are the first work to investigate and propose more robust energy models for hybrid semi-supervised techniques.

- Good Generalisation Capabilities. In contrast to existing techniques that only present results on natural images, we provided an extensive comparison using natural and medical images. Medical images are more complex and fundamentally different than natural images [102], and therefore, our results support the good generalisation capability of our technique. At this point in time, our technique set a new SOTA for semi-supervised techniques.

6 Conclusion

In this work we tackle the problem of classifying with scarce annotations via semi-supervised learning. For this purpose, we proposed a new hybrid framework for semi-supervised classification called CREPE (1-Laplacian gRaph Energy for Pseudo-labEls). In contrast with existing techniques that focus on developing better mechanisms for improving the network performance, we address the problem of how to design better energy models for pseudo-labelling. The highlight of our work is a novel energy model based on the non-smooth _l_1 norm of the normalised graph 1-Laplacian with thoughtfully selected class priors. Unlike existing deep learning or hybrid techniques, we provided a theoretical analysis of our model. We provide a convergence analysis for our model and its properties. We also show that energy models provide better pseudo-labels than the ones directly obtained from a network. We supported our model by an extensive evaluation using major datasets composed of natural and medical images. We showed that our technique is able to provide state-of-the-art performance for semi-supervised classification.

Acknowledgments

AI Aviles-Rivero gratefully acknowledges support from CMIIH and CCIIMI, University of Cambridge. This project has also received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 77782. CB Schönlieb acknowledges support from the Leverhulme Trust project on ‘Breaking the non-convexity barrier’, the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC EP/S026045/1, EP/T003553/1 and EP/N014588/1, the Wellcome Innovator Award RG98755, the CCIIMI and the Alan Turing Institute. RT Tan research in this work is supported by MOE2019-T2-1-130.

References

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” International Conference on Learning Representations (ICLR), 2015.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.
[3] O. Chapelle, B. Scholkopf, and A. Zien, “Semi-supervised learning,” MIT Press, vol. 20, no. 3, pp. 542–542, 2006.
[4] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” Synthesis lectures on artificial intelligence and machine learning, 2009.
[5] X. Zhu, Z. Ghahramani, and J. D. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in International conference on Machine learning (ICML’03), 2005, pp. 912–919.
[6] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, “Learning with local and global consistency,” in Advances in neural information processing systems, 2004, pp. 321–328.
[7] J. Wang, T. Jebara, and S.-F. Chang, “Graph transduction via alternating minimization,” in International conference on Machine learning (ICML). ACM, 2008, pp. 1144–1151.
[8] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” Technical Report CMU-CALD-02-107, Carnegie Mellon University, Tech. Rep., 2002.
[9] T. Joachims, “Transductive learning via spectral graph partitioning,” in International Conference on Machine Learning (ICML), 2003, pp. 280–297.
[10] M. Hein and T. Bühler, “An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca,” in Advances in Neural Information Processing Systems (NIPS), 2010, pp. 847–855.
[11] Y.-M. Zhang, K. Huang, and C.-L. Liu, “Fast and robust graph-based transductive learning via minimum tree cut,” in IEEE International Conference on Data Mining, 2011, pp. 952–961.
[12] X. Zhu and J. Lafferty, “Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning,” in International conference on Machine Learning (ICML), 2005, pp. 1052–1059.
