Hybrid Energy Based Model in the Feature Space for Out-of-Distribution Detection

Marc Lafon [1] Elias Ramzi [1 2] Clément Rambour [1] Nicolas Thome [3]

Abstract

Out-of-distribution (OOD) detection is a critical requirement for the deployment of deep neural networks. This paper introduces the HEAT model, a new post-hoc OOD detection method estimating the density of in-distribution (ID) samples using hybrid energy-based models (EBM) in the feature space of a pre-trained backbone. HEAT complements prior density estimators of the ID density, *e.g.* parametric models like the Gaussian Mixture Model (GMM), to provide an accurate yet robust density estimation. A second contribution is to leverage the EBM framework to provide a unified density estimation and to compose several energy terms. Extensive experiments demonstrate the significance of the two contributions. HEAT sets new state-of-the-art OOD detection results on the CIFAR-10 / CIFAR-100 benchmark as well as on the large-scale Imagenet benchmark. The code is available at: `github.com/MarcLafon/heatood`.

1. Introduction

Out-of-distribution (OOD) detection is a major safety requirement for the deployment of deep learning models in critical applications, *e.g.* healthcare, autonomous steering, or defense (Bendale & Boult, 2015; Amodei et al., 2016; Janai et al., 2020). Deployed machine learning systems must successfully perform a specific task, *e.g.* image classification, or image segmentation while being able to distinguish *in-distribution* (ID) from OOD samples, in order to abstain from making an arbitrary prediction when facing the latter.

OOD detection is a challenge for state-of-the-art deep neural networks. Most recent approaches follow a post-hoc strategy (Hendrycks & Gimpel, 2017; Liang et al., 2018a; Liu et al., 2020; Sehwag et al., 2021; Sun et al., 2022; Wang et al., 2022) suitable for real-world purpose, which offers the possibility to leverage state-of-the-art models for the main prediction task and to maintain their performances. It also relaxes the need for very demanding training processes, which can be prohibitive with huge deep neural nets and foundation models (Bommasani et al., 2021; Radford et al., 2021; Rombach et al., 2022; Alayrac et al., 2022).

Post-hoc methods exploit the feature space of a pre-trained network and attempt at estimating the density of ID features to address OOD detection. Existing ID density estimation methods include Gaussian Mixture Models (GMMs) (Lee et al., 2018b; Sehwag et al., 2021), the nearest neighbors distribution (Sun et al., 2022), or the distribution derived from the energy logits (EL) (Liu et al., 2020). However, these approaches tend to detect different types of OOD data: for instance, GMMs' density explicitly decreases when moving away from training data, making them effective for far-OOD[1] detection, while EL benefits from the classifier training to obtain strong results on near-OOD samples (Wang et al., 2022).

In this work, we introduce **HEAT**, a new density-based OOD detection method which estimate the density of ID samples using a **H**ybrid **E**nergy based model in the fe**AT**ure space of a fixed pre-trained backbone, which provides strong OOD detection performances on both near and far-OOD data. HEAT leverages the energy-based model (EBM) framework (LeCun et al., 2006) to build a powerful density estimation method relying on two main components:

1. **Energy-based correction** of prior OOD detectors (*e.g.* GMMs or EL) with a data-driven EBM, providing an accurate ID density estimation while benefiting from the strong generalization properties of the priors. The corrected model is carefully trained such that the prior and residual terms achieve optimal cooperation.
2. **Hybrid density estimation** grounded by a sound energy functions composition combining several sources to improve OOD detection. The energy composition requires a single hyper-parameter, and involves no computational overhead since it is applied at a single layer of the network.

[1]Cedric Laboratory, Cnam, Paris, France [2]Coexya [3]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France. Correspondence to: Marc Lafon <marc.lafon@lecnam.net>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] We denote as far (resp. near) OOD samples with classes that are semantically distant (resp. close) from the ID classes.

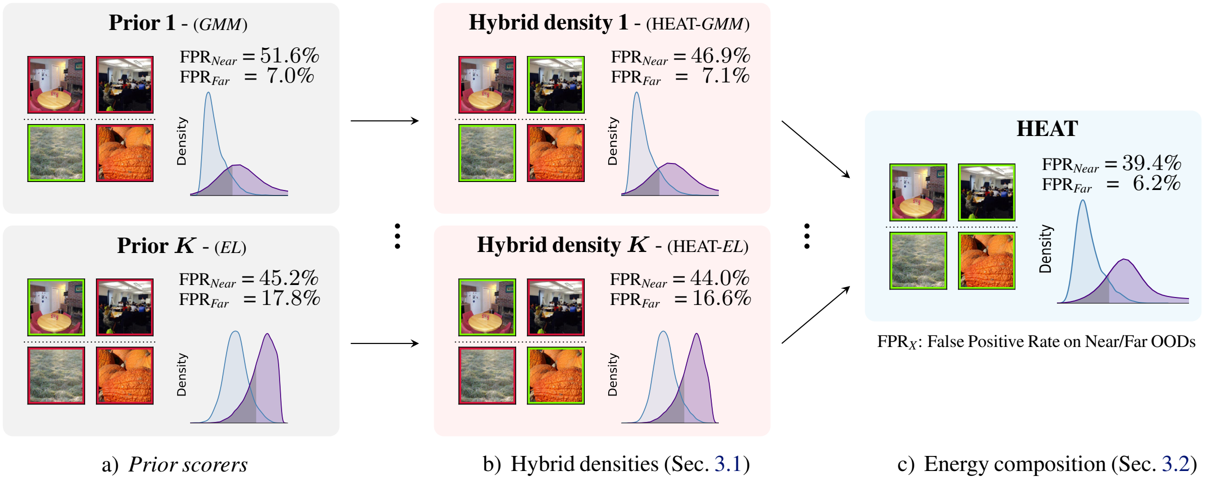


Figure 1. **Illustration of our HEAT model**. HEAT leverages a) K prior density estimators, such as GMM or EL, and overcomes their modeling biases by learning a residual term with an EBM b) leading to more accurate OOD scorers, *e.g.* HEAT-GMM or HEAT-EL. The second contribution is to combine the different refined scorers using an EBM energy composition function. The final HEAT prediction c) can thus leverage the strengths of the different OOD scorers, and be effective for both far and near-OOD detection.

We illustrate HEAT in Fig. 1 using two prior OOD detectors from the literature: SSD+ which is based on GMMs (Sehwag et al., 2021) and EL (Liu et al., 2020), with CIFAR-10 dataset as ID dataset and with six OOD datasets, see Sec. 4. We can see in Fig. 1 that GMM is able to correctly detect far-OOD samples while struggling on near-OOD samples when EL exhibits the opposite behavior. The energy-correction step enhance both priors, reducing the false positive rate (FPR) by -4.7 pts on near-OOD while being stable on far-OOD for GMM, and by -3.2 pts on near-OOD and -1.2 pts for EL. Finally, the energy-composition step produces a hybrid density estimator leading to a better ID density estimation which further improves the OOD detection performances, both for near and far OOD regimes.

We conduct an extensive experimental validation in Sec. 4, showing the importance of our two contributions. HEAT sets new state-of-the-art OOD detection results with CIFAR-10/-100 as ID data, but also on the large-scale Imagenet dataset. HEAT is also agnostic to the prediction backbone (ResNet, ViT) and remains effective in low-data regimes.

2. Related work

Seminal attempts for OOD detection used supervised methods based on external OOD samples (Lee et al., 2018a; Malinin & Gales, 2018) or "Outlier Exposure" (OE) (Hendrycks et al., 2019) enforcing a uniform OOD distribution. Although OOD datasets can improve OOD detection, their relevance is questionable since collecting representative OOD datasets is arguably impossible as OOD lie anywhere outside the training distribution (Charpentier et al., 2020). It can also have the undesirable effect of learning detectors biased towards certain types of OOD (Wang et al., 2022).

Density-based OOD detection. Estimating the density of ID training samples to perform OOD detection is a natural strategy that has been widely explored. In their seminal work, (Lee et al., 2018b) first proposed to approximate the ID features density with a class-conditional GMM. Subsequent works adopted the same approach by adding slight modifications. For instance, (Sehwag et al., 2021) proposed to learn the GMM density of normalized features without having access to class labels. Recently, (Sun et al., 2022) challenged the GMM distributional assumption by showing that using a deep nearest neighbors approach on normalized features has strong OOD detection performances.

Energy-**B**ased **M**odels (**EBM**) are another approach to estimate the ID density which have made incredible progress in generative modeling for images in recent years (Xie et al., 2016; Du & Mordatch, 2019; Grathwohl et al., 2020). However, their performances for OOD detection are not yet comparable with OOD methods based on the feature space (Elflein et al., 2021). (Liu et al., 2020) have proposed to perform OOD detection with an energy score defined by the *logsumexp* of the logits (EL) of the pre-trained classifier showing improvement over using the classifier's predicted probabilities (Hendrycks & Gimpel, 2017). Furthermore, the authors of EL propose to fine-tune the logits of the classifier using external OOD datasets. Contrarily, we do not use any OOD to learn HEAT but rely on proper EBM training to estimate the ID features density.

Energy-based correction. Our method rely on energy-based correction of a reference model. This idea has been explored in noise contrastive estimation (NCE) (Gutmann &

Hyvärinen, 2010) where the correction is obtained by discriminative learning. Learning an EBM in cooperation with a generator model has been introduce in (Xie et al., 2018) where an EBM learns to refine generated samples and has also been applied to cooperative learning of an EBM with a conditional generator (Xie et al., 2022a), a VAE (Pang et al., 2020; Xie et al., 2021; Xiao et al., 2021) a normalizing flow (Nijkamp et al., 2022; Xie et al., 2022b). Contrarily to our method which is designed for OOD detection, previous works focus on generation and cannot benefit from a fixed prior OOD detector as they use a cooperative learning strategy.

Residual learning. Training hybrid models, where a data-driven *residual* complements an approximate predictor, has been proposed in several context, *e.g.* in complex dynamic forecasting (Yin et al., 2021), in NLP (Bakhtin et al., 2021), in video prediction (Le Guen & Thome, 2020; Le Guen et al., 2022), or in robotics (Zeng et al., 2020). Such residual approaches have also emerged for OOD detection. ResFlow (Zisselman & Tamar, 2020) uses a normalizing flow (NF) to learn the residual of a Gaussian density for OOD detection. The approach is related to ours, but NFs require invertible mapping, which intrinsically limit their expressive power and make the learned residual less accurate. Also, ViM (Wang et al., 2022) proposes to model the residual of the ID density by using the complement to a linear manifold on the ID manifold. With HEAT, we can learn a non-linear residual and include a residual from different energy terms to improve ID density modeling. We verify experimentally that HEAT significantly outperforms these two baselines for OOD detection.

Ensembling & composition. The question of merging several networks, also known as *ensembling* (Lakshminarayanan et al., 2017) has been among the first and most successful approaches for OOD detection. The ensemble can include different backbones or different training variants. For OOD detection, several post-hoc approaches also model the ID density at different layer depth of a pre-trained model, the overall density score being obtained by ensembling such predictions (Lee et al., 2018b; Sastry & Oore, 2020; Zisselman & Tamar, 2020). The main limitation of these approaches relates to their computational cost since the inference time is proportional to the number of networks. The overhead quickly becomes prohibitive in contexts with limited resources. Several sources of prior densities are combined in (Wang et al., 2022) to refine OOD detection. Our approach is a general framework adapting the EBM composition model (Du et al., 2020; 2021) to OOD detection, and can thus include several hybrid energy terms to refine ID density estimation. In terms of computational cost, we apply our model at a single layer of the network, bringing essentially no computational overhead at inference (see Appendix B.2).

3. HEAT for OOD detection

In this section, we describe the proposed HEAT model to estimate the density of in-distribution (ID) features using a hybrid energy-based model (EBM). We remind that we place ourselves in the difficult but realistic case where only ID samples are available, and we do not use any OOD samples for density estimation. Also, HEAT is a post-hoc approach estimating the density of the latent space of a pre-trained prediction model, as in (Lee et al., 2018b; Wang et al., 2022; Sehwag et al., 2021; Sun et al., 2022).

Let $p(\mathbf{x})$ be the probability of ID samples, where $\mathbf{x} \in \mathcal{X}$, and $\mathbf{z} = \phi(\mathbf{x}) \in \mathcal{Z}$ denotes the network's embedding of $\mathbf{x}$ with $\mathcal{Z}$ the latent space at the penultimate layer of a pre-trained prediction model f, *e.g.* a deep neural net for classification. We aim at estimating $p(\mathbf{z}|\mathcal{D})$ with $\mathcal{D} := \{\mathbf{x}_i\}_{i=1}^N$ the ID training dataset[2].

We illustrate the two main components at the core of HEAT in Fig. 2. Firstly, we introduce a hybrid density estimation to refine a set of prior densities $\{q_k(z)\}_{1\leq k \leq K}$ by complementing each of them with a residual EBM. Secondly, we propose to compose several hybrid density estimations based on different priors, which capture different facets of ID density distributions.

3.1. Hybrid Energy-based density estimation

The main motivation in hybrid EBM density estimation is to leverage existing models that rely on specific assumptions on the form of the density $p(\mathbf{z})$, *e.g.*EL (Liu et al., 2020), which captures class-specific information in the logit vector, or SSD (Sehwag et al., 2021) which uses a GMM. These approaches have appealing properties: GMM is a parametric model relying on few parameters thus exhibiting strong generalization performances, and EL benefits from classification training. However, their underlying modeling assumptions intrinsically limit their expressiveness which leads to coarse boundaries between ID and OOD, and they generally fail at discriminating between ambiguous data.

Hybrid EBM model. Formally, let $q_k(\mathbf{z})$ be a density estimator inducing an OOD-prior among a set of K priors $\{q_k(\mathbf{z})\}_{1\leq k \leq K}$. We propose to refine its estimated density by learning a residual model $p^r_{\theta_k}(\mathbf{z})$, such that our hybrid density estimation is performed by $p^h_{\theta_k}(\mathbf{z})$ as follows:

$$p^h_{\theta_k}(\mathbf{z}) = \frac{1}{Z(\theta_k)} p^r_{\theta_k}(\mathbf{z}) q_k(\mathbf{z}), \tag{1}$$

with $Z(\theta_k) = \int p^r_{\theta_k}(\mathbf{z}) q_k(\mathbf{z}) \mathrm{d}\mathbf{z}$ the normalization constant. We propose to learn the residual density $p^r_{\theta_k}(\mathbf{z})$ with an EBM: $p^r_{\theta_k}(\mathbf{z}) \propto \exp(-E_{\theta_k}(\mathbf{z}))$. From Eq. (1), we can derive

[2] we ignore the dependence to $\mathcal{D}$ in the following and denote the sought density as $p(\mathbf{z})$.

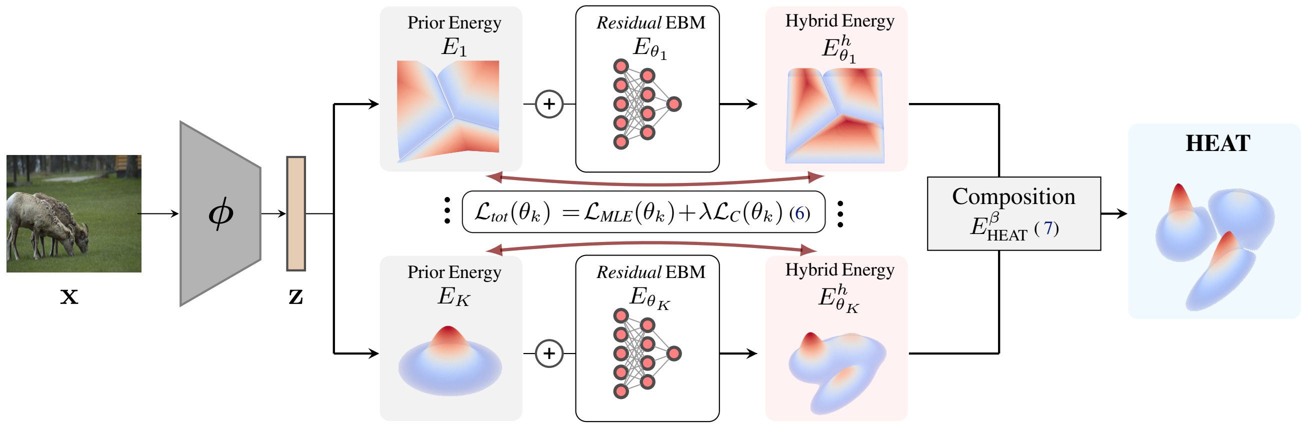


Figure 2. **Schematic view of the HEAT model for OOD detection**. Each selected prior density estimator q_k is expressed as an EBM, $q_k(\boldsymbol{z}) \propto \exp(-E_{q_k}(\boldsymbol{z}))$, and is refined with its own residual EBM parameterized with a neural network: The energy for each prior E_k (*e.g.* EL, GMM) is corrected by a residual energy E_{θ_k} to produce an hybrid energy $E^h_{\theta_k}$ (cf. Sec. 3.1). Then all hybrid energies are composed to produce HEAT's energy E^β_{HEAT} (cf. Sec. 3.2), which is used as uncertainty score for OOD detection.

a hybrid energy $E^h_{\theta_k}(\mathbf{z}) = E_{q_k}(\mathbf{z}) + E_{\theta_k}(\mathbf{z})$ and express $p^h_{\theta_k}(\mathbf{z})$ as follows:

$$p^h_{\theta_k}(\mathbf{z}) = \frac{1}{Z(\theta_k)} \exp\left(-E^h_{\theta_k}(\mathbf{z})\right), \tag{2}$$

with $E_{q_k} = -\log q_k(\mathbf{z})$ the energy from the prior. The goal of the residual energy $E_{\theta_k}(\mathbf{z})$ is to compensate for the lack of accuracy of the energy of the prior density $q_k(\mathbf{z})$. We choose to parameterize it with a neural network, as shown in Fig. 2. This gives our EBM density estimation the required expressive power to approximate the residual term.

Hybrid EBM training. The hybrid model energy $E^h_{\theta_k}(\mathbf{z})$ can be learned via maximum likelihood estimation (MLE), which amounts to perform stochastic gradient descent with the following loss (cf. Appendix A.1 for details):

$$\mathcal{L}_{MLE}(\theta_k) = \mathbb{E}_{\mathbf{z}\sim p_{in}}\left[E_{\theta_k}(\mathbf{z})\right] - \mathbb{E}_{\mathbf{z}'\sim p^h_{\theta_k}}\left[E_{\theta_k}(\mathbf{z}')\right], \tag{3}$$

with $\mathbf{z} \sim p_{in}$ being the true distribution of the features from the dataset. Minimizing Eq. (3) has for effect to lower the energy of real samples while raising the energy of generated ones. To learn a residual model we *must* sample $\mathbf{z}'$ from the hybrid model $p^h_{\theta_k}$. To do so, we follow previous works on EBM training (Du & Mordatch, 2019) and exploit stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011). SGLD sampling consists in gradient descent on the energy function:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \frac{\eta}{2}\nabla_{\mathbf{z}} E^h_{\theta_k}(\mathbf{z}_t) + \sqrt{\eta}\, w_t, \text{ with } w_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{4}$$

where η is the step size, the chain being initiated with $\mathbf{z}_0 \sim q_k$. The residual energy corrects the prior density by raising (resp. lowering) the energies in areas where the prior over- (resp. under-) estimates p_{in}. It then does so for the current hybrid model $E^h_{\theta_k}$. The overall training hybrid EBM scheme is summarized Algorithm 1.

Controlling the residual. As our goal is to learn a residual model over q, we must prevent the energy-correction term E_{θ_k} to take too large values thus canceling the benefit from the prior model q_k. Therefore, we introduce an additional loss term preventing the hybrid model from deviating to much from the prior density:

$$\mathcal{L}_C(\theta_k) = \mathbb{E}_{p_{in}, p^h_{\theta_k}}\left[\,(E^h_{\theta_k} - E_{q_k})^2\,\right]. \tag{5}$$

The final loss is then:

$$\mathcal{L}_{Tot}(\theta_k) = \mathcal{L}_{MLE}(\theta_k) + \lambda\, \mathcal{L}_C(\theta_k), \tag{6}$$

where λ is an hyper-parameter balancing between the two losses. Although $\mathcal{L}_C(\theta_k)$ in Eq. (5) rewrites as $\mathbb{E}_{p_{in}, p^h_{\theta_k}}\left[E^2_{\theta_k}\right]$, we point out that its objective goes beyond a standard ℓ_2-regularization used to stabilize training. It has the more fundamental role of balancing the prior and the residual energy terms in order to drive a proper cooperation.

3.2. Composition of refined prior density estimators

In this section, we motivate the choice of prior OOD scorers that we correct, and how to efficiently compose them within our HEAT framework.

Selected OOD-Priors. As previously stated, EL and GMM show complementary OOD detection performances, EL being useful to discriminate class ambiguities while GMM is effective on far-OOD. Additionally they can be directly interpreted as energy-based models and thus can easily be

Algorithm 1 Hybrid Energy Based Model Training

```
input  : Features $\mathcal{D}_z$, ID-Prior $(q_k, E_{q_k})$, $\lambda$, $\alpha$ and $\eta$.
output : Hybrid EBM $E^h_{\theta_k} = E_{q_k}(\boldsymbol{z}) + E_{\theta_k}(\boldsymbol{z})$.   // cf. Eq. (2)
while not converged do
    Sample $\boldsymbol{z} \in \mathcal{D}_z$ and $\boldsymbol{z}'_0 \sim q_k$
    for $0 \leq t \leq T-1$ do
        $w \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
        $\boldsymbol{z}'_{t+1} \leftarrow \boldsymbol{z}_{t'} - \frac{\eta}{2} \nabla_{\boldsymbol{z}} E^h_{\theta_k}(\boldsymbol{z}'_t) + \sqrt{\eta} w$   // SGLD, Eq. (4)
    end
    $\mathcal{L}_{Tot}(\theta_k) = \mathcal{L}_{MLE}(\theta_k) + \lambda \mathcal{L}_c(\theta_k)$   // cf. Eq. (6)
    $\theta_k \leftarrow \theta_k - \alpha \nabla_{\theta_k} \mathcal{L}_{Tot}(\theta_k)$
end
```

refined and composed with HEAT. Based on the energy from the logits derived in (Liu et al., 2020) we express the hybrid energy HEAT-EL as $E^h_{\theta_l}(\mathbf{z}) = \text{-}\log \sum_c e^{f(\mathbf{z})[c]} + E^r_{\theta_l}(\mathbf{z})$ where $f(.)[c]$ denotes the logit associated to the class c.

For the GMM prior, we derive an energy from the Mahalanobis distances to each class centroid. Giving the following expression for our hybrid HEAT-GMM's energy $E^h_{\theta_g}(\mathbf{z}) = -\log \sum_c e^{-\frac{1}{2}(\mathbf{z}-\mu_c)^T \mathbf{\Sigma}^{-1}(\mathbf{z}-\mu_c)} + E^r_{\theta_g}(\mathbf{z})$ with Σ and μ_c being the empirical covariance matrix and mean feature for class c. HEAT-GMM's energy is computed on the $\mathbf{z}$ vector in Fig. 2, which is obtained by average pooling from the preceding tensor in the network. To improve HEAT's OOD detection performances, we propose to further exploit feature volume prior to the pooling operation (*e.g.* average pooling) as we hypothesize that it contains more information relevant to OOD detection. To do so, we compute the vector of second-order moments of the feature volume by using a std-pooling operator and subsequently model the density of the second-order features with a GMM. This leads to a third hybrid EBM denoted as HEAT-GMM$_{std}$.

Note that our HEAT method can be extended to K prior scorers, provided that they can write as an EBM and that they are differentiable in order to perform SGLD sampling. Interesting extensions would include adapting the approach to other state-of-the-art OOD detectors, such as a soft-KNN (Sun et al., 2022) or ViM (Wang et al., 2022). We leave these non-trivial extensions for future works.

Composition strategy. The EBM framework offers a principled way to make a composition (Du et al., 2020) of energy functions. Given K corrected energy functions $E^h_{\theta_k}$, such that: $p^h_{\theta_k} \propto \exp(-(E_{\theta_k}(\mathbf{z}) + E_{q_k}(\mathbf{z}))$, we introduce the following composition function:

$$E^{\beta}_{\text{HEAT}} = \frac{1}{\beta} \log \sum_{k=1}^{K} e^{\beta E^h_{\theta_k}} \tag{7}$$

Depending on β, E^{β}_{HEAT} can recover a sum of energies ($\beta = 0$), *i.e.* a product of probabilities. For $\beta = \text{-}1$, E^{β}_{HEAT} is equivalent to the *logsumexp* operator, *i.e.* a sum of probabilities. More details in Appendix A.2. Moreover, unlike previous approaches that require learning a set of weights (Lee et al., 2018b; Zisselman & Tamar, 2020), HEAT's composition only requires tuning a single hyper-parameter, *i.e.* β which has a clear interpretation.

The composition strategy adopted in HEAT is also scalable since: i) we work in the feature space $\mathbf{z} = \phi(\mathbf{x}) \in \mathcal{Z}$ of controlled dimension (*e.g.* 1024 even for the CLIP foundation model (Radford et al., 2021)), and ii) our energy-based correction uses a relatively small model (we use a 6-layers MLP in practice). We study the computational cost of HEAT in Appendix B.2 and show the large gain in efficiency compared to *e.g.* deep ensembles.

OOD detection with HEAT. Finally, we use the learned and composed energy of HEAT, E^{β}_{HEAT} in Eq. (7), as an uncertainty score to detect OOD samples.

4. Experiments

Datasets. We validate HEAT on several benchmarks. The two commonly used CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) benchmarks as in (Sehwag et al., 2021; Sun et al., 2022). We also conduct experiments on the large-scale Imagenet (Deng et al., 2009) dataset. More details in Appendix B.

Evaluation metrics. We report the following standard metrics used in the literature (Hendrycks & Gimpel, 2017): the area under the receiver operating characteristic curve (AUC) and the false positive rate at a threshold corresponding to a true positive rate of 95% (FPR95).

Implementation details. All results on CIFAR-10 and CIFAR-100 are reported using a ResNet-34 (He et al., 2016), on Imagenet we use the pre-trained ResNet-50 from `PyTorch` (Paszke et al., 2019). We detail all implementation details in Appendix B.

Baselines. We perform extensive validation of HEAT *vs.* several recent state-of-the-art baselines, including the maximum softmax probability (MSP) (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018b), Energy-logits (Liu et al., 2020), SSD (Sehwag et al., 2021), KNN (Sun et al., 2022) and ViM (Wang et al., 2022). We apply our energy-based correction of EL, GMM and GMM$_{std}$ that we then denote as HEAT-EL, HEAT-GMM and HEAT-GMM$_{std}$. We choose those priors as they can naturally be written as energy models as described in Sec. 3.1, furthermore, they are strong baselines and combining them allows us to take advantage of their respective strengths (discussed in Sec. 3.2). All the baselines are compared using the same backbone trained with the standard cross-entropy loss.

Table 1. Refinement of Energy-logits (Liu et al., 2020) (EL) and GMM, GMM with std-pooling (GMM$_{std}$) with our energy-based correction on CIFAR-10 and CIFAR-100 as in-distribution datasets. Results are reported with FPR95↓ / AUC ↑.

	Method	*Near-OOD*		*Mid-OOD*		*Far-OOD*		Average
		C-100/10	**TinyIN**	**LSUN**	**Places**	**Textures**	**SVHN**	
CIFAR-10	EL	48.4 / 86.9	41.9 / 88.2	33.7 / 92.6	35.7 / 91.0	30.7 / 92.9	4.9 / 99.0	32.6 / 91.8
	HEAT-EL	**47.3 / 88.0**	**40.7 / 88.9**	**30.8 / 93.4**	**33.8 / 91.8**	**28.8 / 93.9**	**4.5 / 99.1**	**31.0 / 92.5**
	GMM	52.6 / 89.0	50.9 / 89.5	47.1 / 92.4	46.4 / 91.2	**13.1 / 97.8**	**0.9 / 99.8**	35.1 / 93.3
	HEAT-GMM	**49.0 / 89.8**	**44.8 / 90.4**	**40.5 / 93.2**	**40.4 / 92.0**	13.4 / 97.7	**0.8 / 99.8**	**31.5 / 93.8**
	GMM$_{std}$	58.4 / 84.9	50.6 / 87.9	32.2 / 94.5	38.5 / 91.8	13.8 / **97.6**	**2.5 / 99.5**	32.7 / 92.7
	HEAT-GMM$_{std}$	**56.1 / 86.1**	**47.8 / 88.7**	**28.2 / 95.2**	**35.8 / 92.5**	**13.3 / 97.5**	2.7 / 99.4	**30.7 / 93.2**
CIFAR-100	EL	80.6 / 76.9	79.4 / 76.5	87.6 / 71.7	83.1 / 74.7	62.4 / 85.2	53.0 / 88.9	74.3 / 79.0
	HEAT-EL	**80.1 / 77.2**	**77.6 / 77.5**	**87.2 / 72.2**	**81.8 / 75.0**	**61.5 / 85.8**	**47.5 / 90.2**	**72.6 / 79.6**
	GMM	85.6 / 73.6	82.5 / 77.2	87.8 / 73.7	84.5 / 74.4	**36.7 / 92.4**	20.0 / 96.3	66.2 / 81.3
	HEAT-GMM	**84.2 / 74.8**	**80.5 / 78.5**	**86.4 / 74.8**	**82.7 / 75.9**	37.9 / **92.2**	**17.8 / 96.7**	**64.9 / 82.1**
	GMM$_{std}$	91.4 / 67.9	84.3 / 74.8	83.4 / 75.2	83.5 / 75.2	**40.6 / 91.3**	36.7 / 93.1	70.0 / 79.6
	HEAT-GMM$_{std}$	**89.1 / 70.3**	**82.2 / 76.2**	**82.3 / 76.1**	**81.4 / 76.7**	42.9 / 90.7	**32.9 / 93.8**	**68.5 / 80.6**

4.1. HEAT improvements

In this section we study the different components of HEAT. In Tab. 1 we show that learning a residual correction term with HEAT improves the OOD detection performances of prior scorers. In Tab. 2 we show the interest of learning a residual model as described in Sec. 3.1 rather than a standard fully data-driven energy-based model. Finally in Tab. 3 we show how using the energy composition improves OOD detection.

Correcting prior scorers. In Tab. 1 we demonstrate the effectiveness of energy-based correction to improve different prior OOD scorers on two ID dataset: CIFAR-10 and CIFAR-100. We show that across the two ID datasets and for all prior scorers, using a residual corrections always improves the aggregated results, *e.g.* for GMM -3.6 pts FPR95 on CIFAR-10 and -1.3 pts FPR95 on CIFAR-100. Furthermore on near-OOD and mid-OOD learning our correction always improves the prior scores, *e.g.* on LSUN with CIFAR-10 as ID dataset the correction improves EL by -2.9 pts FPR95, GMM by -6.6 pts FPR95 and -4 pts FPR95 for GMM$_{std}$. On far-OOD the corrected scorers performs at least on par with the base scorers, and can further improve it, *e.g.* on SVHN when CIFAR-100 is the ID datasets, the correction improves by -5.5pts FPR95, -2.2 pts FPR95 and -3.8pts FPR95, EL, GMM, GMM$_{std}$ respectively. Overall Tab. 1 clearly validates the relevance of correcting the modeling assumptions of prior scorers with our learned energy-based residual.

Learning a residual model. In Tab. 2 we compare learning an EBM (cf. Appendix A.1) *vs.* our residual training using a GMM prior (HEAT-GMM) of Sec. 3 on CIFAR-10 and CIFAR-100. The EBM is a fully data-driven approach, which learns the density of ID samples without any prior distribution model. On both datasets, our residual training leads to better performances than the EBM, *e.g.* +2.6 pts AUC on CIFAR-100. On near-OOD, both the residual training and the EBM perform on par. On far-OOD, our residual training takes advantage of the good performances of the prior scorer, *i.e.*GMM, and significantly outperforms the EBM, especially on CIFAR-100, with *e.g.* +7.7 pts AUC on Textures. Our residual training combines the strengths of GMM and EBMs: Gaussian modelization by design penalizes samples far away from the training dataset and thus eases far-OOD's detection, whereas EBM may overfit in this case. On the other hand, near-OOD detection requires a too complex density estimation for simple parametric distribution models such as GMMs.

Table 2. Comparison of learning a residual model, *i.e.* HEAT-GMM, *vs.* learning an EBM and GMM. Results reported with AUC ↑.

	Method	*Near-OOD*		*Mid-OOD*		*Far-OOD*		Average
		C-100/10	**TinyIN**	**LSUN**	**Places**	**Textures**	**SVHN**	
C-10	GMM	89.0	89.5	92.4	91.2	**97.7**	**99.8**	93.3
	EBM	89.4	89.9	**93.8**	91.8	96.2	99.0	93.3
	HEAT-GMM	**89.8**	**90.4**	93.2	**92.0**	**97.7**	**99.8**	**93.8**
C-100	GMM	73.6	77.0	73.8	74.5	**92.4**	96.4	81.3
	EBM	**74.8**	**79.7**	71.9	75.4	84.5	91.0	79.5
	HEAT-GMM	**74.8**	78.5	**74.8**	**75.9**	92.2	**96.7**	**82.1**

Table 3. Aggregated performances on CIFAR-10 and CIFAR-100 for the energy composition of the refined OOD scorers of Tab. 1.

HEAT -GMM	HEAT -GMM$_{std}$	HEAT -EL	**CIFAR-10**		**CIFAR-100**	
			FPR95↓	AUC ↑	FPR95↓	AUC ↑
✓	✗	✗	31.5	93.8	64.9	82.1
✗	✓	✗	30.7	93.2	68.5	80.6
✗	✗	✓	31.0	92.5	72.6	79.6
✓	✓	✗	25.6	94.6	64.3	82.7
✓	✗	✓	28.0	94.1	65.5	82.4
✗	✓	✓	23.6	94.6	66.6	82.1
✓	✓	✓	**23.5**	**94.8**	**63.9**	**83.0**

Table 4. **Results on CIFAR-10 & CIFAR-100.** All methods are based on a pre-trained ResNet-34 trained on the ID dataset only. ↑ indicates larger is better and ↓ the opposite. Best results are in bold, second best underlined. Results are reported with FPR95↓ / AUC ↑.

	Method	*Near-OOD*		*Mid-OOD*		*Far-OOD*		**Average**
		C-10/C-100	**TinyIN**	**LSUN**	**Places**	**Textures**	**SVHN**	
CIFAR-10	MSP (Hendrycks & Gimpel, 2017)	58.0 / 87.9	55.9 / 88.2	50.5 / 91.9	52.7 / 90.2	52.3 / 91.7	19.7 / 97.0	48.2 / 91.2
	ODIN (Liang et al., 2018b)	48.4 / 86.0	42.2 / 87.3	32.6 / 92.3	35.6 / 90.4	29.4 / 92.6	7.8 / 98.3	32.6 / 91.1
	KNN (Sun et al., 2022)	47.9 / **90.3**	43.1 / 90.6	36.1 / 94.1	37.9 / 92.7	24.9 / 96.0	8.1 / 98.6	33.0 / 93.7
	ViM (Wang et al., 2022)	44.8 / 89.2	40.1 / 89.8	32.0 / 93.8	34.3 / 92.2	17.9 / 96.4	3.6 / 99.2	28.8 / 93.4
	SSD+ (Sehwag et al., 2021)	52.6 / 89.0	50.9 / 89.5	47.1 / 92.4	46.4 / 91.2	13.1 / **97.8**	**0.9 / 99.8**	35.1 / 93.3
	EL (Liu et al., 2020)	48.4 / 86.9	41.9 / 88.2	33.7 / 92.6	35.7 / 91.0	30.7 / 92.9	4.9 / 99.0	32.6 / 91.8
	DICE (Sun & Li, 2022)	51.0 / 85.7	44.3 / 87.0	33.3 / 92.3	35.6 / 90.5	29.3 / 92.8	3.6 / 99.2	32.8 / 91.3
	HEAT (ours)	**43.1 / 90.2**	**35.7 / 91.3**	**22.2 / 95.8**	**27.4 / 93.9**	**11.3 / 97.9**	**1.1 / 99.8**	**23.5 / 94.8**
CIFAR-100	MSP (Hendrycks & Gimpel, 2017)	**80.0 / 76.6**	78.3 / 77.6	**83.5** / 74.7	81.0 / 76.4	72.1 / 81.0	62.0 / 86.4	76.1 / 78.8
	ODIN (Liang et al., 2018b)	81.4 / 76.4	78.7 / 76.2	86.1 / 72.0	82.6 / 74.5	62.4 / 85.2	80.7 / 80.4	78.6 / 77.5
	KNN (Sun et al., 2022)	82.1 / 74.5	**76.7 / 80.2**	90.1 / 74.4	83.2 / 75.5	47.2 / 90.2	35.6 / 93.6	69.2 / 81.4
	ViM (Wang et al., 2022)	85.8 / 74.3	77.5 / 79.6	86.2 / 75.3	**79.8 / 77.6**	42.3 / 91.9	41.3 / 93.2	68.8 / 82.0
	SSD+ (Sehwag et al., 2021)	85.6 / 73.6	82.5 / 77.2	87.8 / 73.7	84.5 / 74.4	**36.7** / 92.4	**20.0 / 96.3**	66.2 / 81.3
	EL (Liu et al., 2020)	**80.6 / 76.9**	79.4 / 76.5	87.6 / 71.7	83.1 / 74.7	62.4 / 85.2	53.0 / 88.9	74.3 / 79.0
	DICE (Sun & Li, 2022)	81.2 / 75.8	82.4 / 74.2	87.8 / 70.4	84.5 / 73.1	63.0 / 83.8	51.9 / 88.1	75.2 / 77.6
	HEAT (ours)	83.7 / 75.8	77.7 / 79.5	**83.4 / 76.3**	**80.0 / 77.8**	37.1 / **92.7**	21.7 / **96.0**	**63.9 / 83.0**

Composing energy-based scorers. In Tab. 3 we show that composing different energy-based scores (see Sec. 3.2), *i.e.* the selected OOD prior scorers with our energy-based correction as described in Sec. 3.1, improves overall performances on CIFAR-10 and CIFAR-100. For instance composing our HEAT-GMM and HEAT-GMM$_{std}$ leads to improvements of all reported results, *i.e.* on CIFAR-10 -5.1 pts FPR95 and +0.8 pts AUC and on CIFAR-100 -0.6 pts FPR95 and +0.6 pts AUC. Composing the three prior scorers leads to the best results, improving over the best single scorer performances by great margins on CIFAR-10 with -7.1 pts FPR95 and +1 AUC and with smaller margins on CIFAR-100 -0.8 pts FPR95 and +1.1 pts AUC on CIFAR-100. This shows the interest of composing different scorers as they detect different types of OOD. Note that while the composition has the best performances our correction model (HEAT-GMM) already has competitive performances on CIFAR-10 and better performances on CIFAR-100 than state-of-the-art methods reported in Tab. 4.

4.2. Comparison to state-of-the-art

In this section, we present the results of HEAT *vs.* state-of-the-art methods. In Tab. 4 we present our results with CIFAR-10, and CIFAR-100 as ID data, and in Tab. 5 we present our results on the large and complex Imagenet dataset.

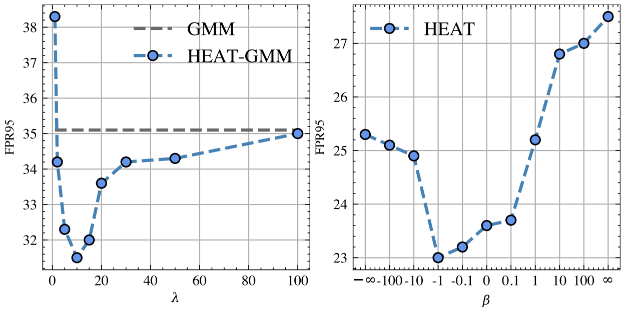
CIFAR-10 results. In Tab. 4 we compare HEAT *vs.* state-of-the-art methods when using CIFAR-10 as the ID dataset. First, we show that HEAT sets a new state-of-the-art on the aggregated results. It outperforms the prior scorers it corrects, *i.e.*SSD+ by -11.6 pts FPR95 and Energy-logits by -9.1 pts FPR95. It also outperforms the previous state-of-the-art methods ViM by -5.3 pts FPR95 and KNN by +1.1 pts AUC. Interestingly we can see that HEAT outperforms other methods because it improves OOD detection on near-, mid-, and far-OOD. On near OOD, it outperforms KNN by -4.6 pts FPR95 on C-100 and Energy-logits by -6.1 pts FPR95 on TinyIN. On mid-OOD detection, it outperforms ViM by -9.8 pts FPR95 on LSUN and Energy-logits by -8.5 pts FPR95. Finally, on far-OOD, the performances are similar to SSD+ which is by far the best performing method on this regime.

CIFAR-100 results. In Tab. 4 we compare HEAT *vs.* state-of-the-art method when using CIFAR-100 as the ID dataset. HEAT outperforms state-of-the-art methods on aggregated results, with -2.3 pts FPR95 and +1.7 pts AUC *vs.*SSD+. HEAT takes advantage of SSD+ on far-OOD and outperforms other methods (except SSD+) by large margins -13.9 pts FPR95 and +2.4 pts AUC on SVHN *vs.* the best non-parametric data-driven density estimation, *i.e.* KNN. Also, HEAT significantly outperforms SSD+ for near-OOD and mid-OOD, *e.g.*-4.8 pts FPR95 on TinyIN or -4.5 pts FPR95 on Places.

Imagenet results. In Tab. 5 we compare HEAT on the recently introduced (Sun et al., 2022) Imagenet OOD benchmark. HEAT sets a new state-of-the-art on this Imagenet benchmark for the aggregated results, with 34.4 FPR95 and 92.6 AUC which outperforms by -1.5 pts FPR95 and +1.7 pts AUC *vs.* the previous best performing method DICE. Furthermore, HEAT improves the aggregated results because it is a competitive method on each dataset. On far-OOD, *i.e.* Textures, it performs on par with SSD+, *i.e.* 5.7 FPR95, the best performing method on this dataset. On mid-OOD, it is the second best method on SUN and on Places behind DICE. Finally, on near-OOD it performs on par with DICE. This shows that HEAT can be jointly effective on

Table 5. **Results on Imagenet.** All methods use an Imagenet pre-trained ResNet-50. Results are reported with FPR95↓ / AUC ↑.

Method	iNaturalist	SUN	Places	Textures	Average
MSP (Hendrycks & Gimpel, 2017)	52.8 / 88.4	69.1 / 81.6	72.1 / 80.5	66.2 / 80.4	65.1 / 82.7
ODIN (Liang et al., 2018b)	41.1 / 92.3	56.4 / 86.8	64.2 / 84.0	46.5 / 87.9	52.1 / 87.8
ViM (Wang et al., 2022)	47.4 / 92.3	62.3 / 86.4	68.6 / 83.3	15.2 / 96.3	48.4 / 89.6
KNN (Sun et al., 2022)	60.0 / 86.2	70.3 / 80.5	78.6 / 74.8	11.1 / 97.4	55.0 / 84.7
SSD+ (Sehwag et al., 2021)	50.0 / 90.7	66.5 / 83.9	76.5 / 78.7	**5.8 / 98.8**	49.7 / 88.0
EL (Liu et al., 2020)	53.7 / 90.6	58.8 / 86.6	66.0 / 84.0	52.4 / 86.7	57.7 / 87.0
DICE (Sun & Li, 2022)	**26.6** / 94.5	**36.5 / 90.8**	**47.9 / 87.5**	32.6 / 90.4	35.9 / 90.9
HEAT (ours)	28.1 / **94.9**	44.6 / **90.7**	58.8 / 86.3	**5.9 / 98.7**	**34.4 / 92.6**



(a) λ *vs.*FPR95↓

(b) β *vs.*FPR95↓

Figure 3. On CIFAR-10 ID: (a) impact of λ in Eq. (6) *vs.* FPR95 and (b) analysis of β in Eq. (7) *vs.* FPR95.

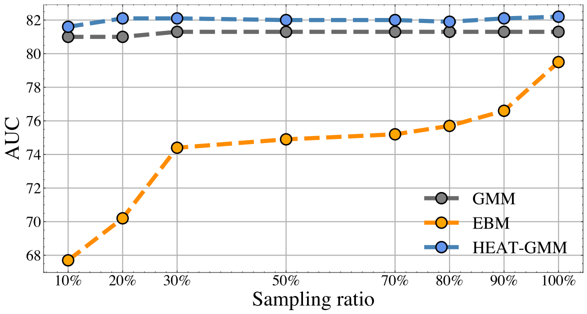


Figure 4. Impact on performances (AUC↑ on CIFAR-100) *vs.* the number of training data for GMM density, fully data-driven EBM, and HEAT. Our hybrid approach maintains strong performances in low-data regime, in contrast to the fully data-driven EBM.

far-, mid-, and near-OOD detection, whereas state-of-the-art methods are competitive for a specific type of OOD only. For instance the performance of DICE drops significantly on Textures. Furthermore, we show in Appendix B.1.3 that using an energy refined version of DICE instead of EL into HEAT's composition further improve OOD detection results. This also shows that HEAT performs well on larger scale and more complex datasets such as Imagenet. In Appendix B.1.1 we show the results of HEAT on the more recent Imagenet OpenOOD benchmarks (Yang et al., 2022), where we show that HEAT also outperforms state-of-the-art methods. In Appendix B.1.4 we show that HEAT also outperforms other methods when using a supervised contrastive backbone. Finally in Appendix B.1.2 we show that HEAT is also state-of-the-art when using another type of neural network, *i.e.* Vision Transformer (Dosovitskiy et al., 2020).

4.3. Model analysis

In this section we show how HEAT works in a wide range of settings. We show in Fig. 3 the impact of λ and β and in Fig. 4 that HEAT performs well in low data regimes.

Robustness to $\boldsymbol{\lambda}$. We show in Fig. 3a the impact of λ on the FPR95 for CIFAR-10 as the ID dataset. We can observe that for a wide range of λ, *e.g.* $[2, 50]$, our energy-based correction improves the OOD detection of the prior scorer, *i.e.*GMM, with ideal values close to ~ 10. λ controls the cooperation between the prior scorer and the learned residual term which can be observed on Fig. 3a. When setting λ to a value that is too low there is no control over the energy. The prior density is completely disregarded which will eventually lead to optimization issues resulting in poor detection performances. On the other hand, setting λ to a value too high (*e.g.* 100) will constrain the energy too much, resulting in performances closer to that of GMM. On CIFAR-10 as the ID dataset we observe similar trends in Appendix B.2.

Robustness to $\boldsymbol{\beta}$. We show in Fig. 3b that HEAT is robust wrt. β in Eq. (7). We remind that $\beta \to 0$ is equivalent to the mean, $\beta \to -\infty$ is equivalent to the minimum and $\beta \to \infty$ is equivalent to the maximum. We show that HEAT is stable to different values of β, and performs best with values close to 0, this is also the case in Appendix B.2. Note that we used $\beta = 0$ for HEAT in Tab. 5 and Tab. 4 but using a lower value, *i.e.*-1, leads to better results. We hypothesize that using a more advanced $\boldsymbol{\beta}$ selection methods could further improve performances.

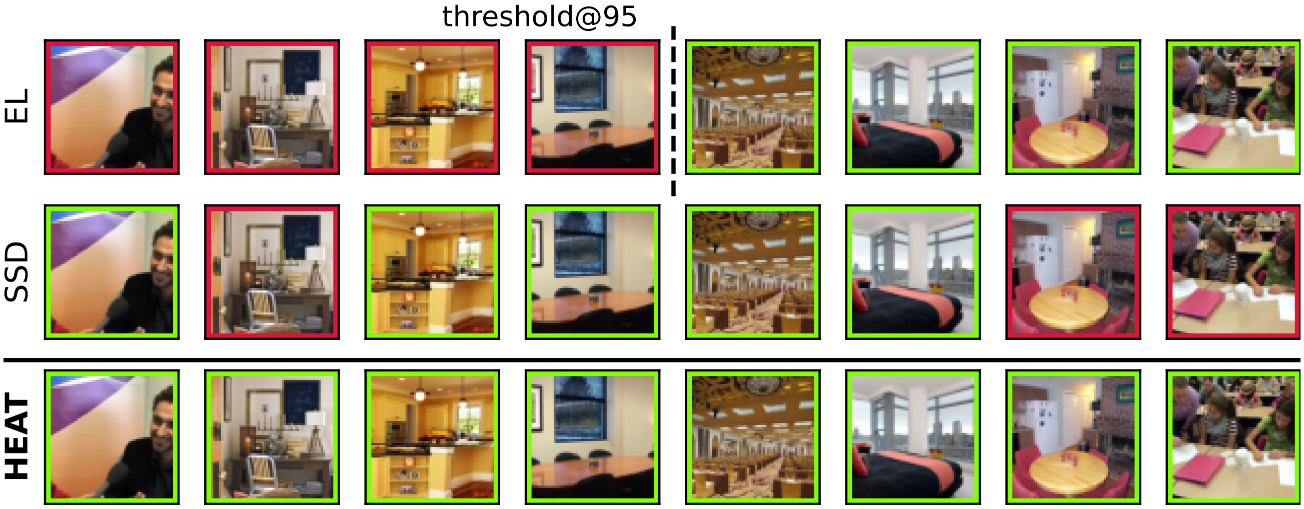


Figure 5. Qualitative comparison of HEAT *vs.*EL (Liu et al., 2020) and SSD (Sehwag et al., 2021). Samples in green are correctly detected as OOD (LSUN), samples in red are incorrectly predicted as ID (CIFAR-10).

Low data regime. We study in Fig. 4 (and Appendix B.2) the stability of HEAT on low data regimes. Specifically, we restrict the training of HEAT to a subset of the ID dataset, *i.e.* CIFAR-100. We compare HEAT to a fully data-driven EBM and to a GMM. The EBM is very sensitive to the lack of training data, with a gap of 12 pts AUC between 10% of data and 100%. On the other hand, GMM is quite robust to low data regimes with a minor gap of 0.3 pts AUC between 10% and 100%. HEAT builds on this stability and is able to improve the performance of GMM for all tested sampling ratios. HEAT is very stable to low data regimes which makes it easier to use than a standard EBM, it is also able to improve GMM even when few training data are available.

4.4. Qualitative results

In Fig. 5 we display qualitative results on CIFAR-10 (ID) and show the detection results of OOD samples from LSUN. We display in red OOD samples incorrectly identified as ID samples, *i.e.* below the threshold at 95% of ID samples, and in green OOD samples correctly detected, *i.e.* are above the 95% threshold. We can see that SSD detects different OOD than EL. HEAT correctly predicts all OOD samples. Other qualitative results are provided in Appendix B.3.

5. Conclusion

We have introduced the HEAT model which leverages the versatility of the EBM framework to provide a strong OOD detection method jointly effective on both far and near-OODs. HEAT i) corrects prior OOD detectors to boost their detection performances and ii) naturally combines the corrected detectors to take advantage of their strengths. We perform extensive experiments to validate HEAT on several benchmarks, highlighting the importance of the correction and the composition, and showing that HEAT sets new state-of-the-art performances on CIFAR-10, CIFAR-100, and on the large-scale Imagenet dataset. HEAT is also applicable to different backbones, and remains efficient in low-data regimes.

Future works include extending HEAT by correcting other prior density estimators, *e.g.*KNN. Another interesting direction is to validate HEAT on other tasks, *e.g.* segmentation, or modalities, *e.g.* NLP.

Acknowledgements

This work was done under grants from the DIAMELEX ANR program (ANR-20-CE45-0026) and the AHEAD ANR program (ANR-20-THIA-0002). It was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012645R1 made by GENCI.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, number NeurIPS, 2022. URL http://arxiv.org/abs/2204.14198. 1

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety. pp. 1–29, 2016. URL http://arxiv.org/abs/1606.06565. 1

Bakhtin, A., Deng, Y., Gross, S., Ott, M., Ranzato, M. A., and Szlam, A. Residual Energy-based Models for Text. *Journal of Machine Learning Research*, 22:1–18, 2021. ISSN 15337928. 3

Bendale, A. and Boult, T. E. Towards open world recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1893–1902. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298799. URL https://doi.org/10.1109/CVPR.2015.7298799. 1

Bendale, A. and Boult, T. E. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1563–1572. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.173. URL https://doi.org/10.1109/CVPR.2016.173. 15

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems*, 2020. 2

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 15

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 5, 15

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8, 17

Du, Y. and Mordatch, I. Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://sites.google.com/view/igebm. 2, 4, 14, 15

Du, Y., Li, S., and Mordatch, I. Compositional Visual Generation with Energy Based Models. *Advances in Neural Information Processing Systems*, 2020-Decem, 2020. ISSN 10495258. URL https://energy-based-model.github.io/. 3, 5

Du, Y., Li, S., Sharma, Y., Tenenbaum, J. B., and Mordatch, I. Unsupervised Learning of Compositional Energy Concepts. *Advances in Neural Information Processing Systems*, 2021. URL https://energy-based-model.github.io/comet/http://arxiv.org/abs/2111.03042. 3

Elflein, S., Charpentier, B., Zügner, D., and Günnemann, S. On Out-of-distribution Detection with Energy-based Models. In *ICML Workshop*, 2021. URL https://github.com/selflein/. 2

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. In *8th International Conference on Learning Representations*, 2020. URL http://arxiv.org/abs/1912.03263. 2, 14, 15

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterington, D. M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 297–304. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/gutmann10a.html. 2

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 5

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017. 1, 2, 5, 7, 8

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep Anomaly Detection with Outlier Exposure. In *ICLR*,

2019. URL https://github.com/hendrycks/outlier-exposure. 2

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8759–8773. PMLR, 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html. 15

Hinton, G. E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 1800(14): 1771–1800, 2002. 14

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 677–689, 2021. 15

Janai, J., Güney, F., Behl, A., and Geiger, A. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Found. Trends Comput. Graph. Vis.*, 12 (1-3):1–308, 2020. doi: 10.1561/0600000079. URL https://doi.org/10.1561/0600000079. 1

Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 5, 15

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 3, 18

Le Guen, V. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

Le Guen, V., Rambour, C., and Thome, N. Complementing brightness constancy with deep networks for optical flow prediction. In *European Conference on Computer Vision (ECCV)*. 2022. 3

LeCun, Y., Chopra, S., Hadsell, R., and Huang, F. J. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006. 1

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018a. 2

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018b. 1, 2, 3, 5, 15

Liang, S., Li, Y., and Srikant, R. Enhancing The Reliability of Out-Of-Distribution Image Detection in Neural Networks. In *ICLR*, 2018a. URL https://github.com/facebookresearch/odin. 1

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of International Conference on Learning Representations*, 2018b. 5, 7, 8

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, 2020. URL https://github.com/wetliu/energy_ood. 1, 2, 3, 5, 6, 7, 8, 9, 19

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018. 2

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010. 14

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 15

Nijkamp, E., Gao, R., Sountsov, P., Vasudevan, S., Pang, B., Zhu, S., and Wu, Y. N. MCMC should mix: Learning energy-based model with neural transport latent space MCMC. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=4C93Qvn-tz. 3

Pang, B., Han, T., Nijkamp, E., Zhu, S. C., and Wu, Y. N. Learning latent space energy-based prior model. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020. 3

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural*

Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., 2019. 5, 15

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. 2021. URL http://arxiv.org/abs/2103.00020. 1, 5, 18

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. 1

Sastry, C. S. and Oore, S. Detecting out-of-distribution examples with gram matrices. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:8449–8459, 2020. URL https://github.com/. 3, 15

Sehwag, V., Chiang, M., and Mittal, P. SSD: A unified framework for self-supervised outlier detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=v5gjXpmR8J. 1, 2, 3, 5, 7, 8, 9, 15, 19

Song, Y. and Kingma, D. P. How to train your energy-based models. *CoRR*, abs/2101.03288, 2021. URL https://arxiv.org/abs/2101.03288. 14

Sun, Y. and Li, Y. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. 7, 8

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022. URL https://proceedings.mlr.press/v162/sun22d.html. 1, 2, 3, 5, 7, 8, 15

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. *Proceedings of the 25th International Conference on Machine Learning*, pp. 1064–1071, 2008. doi: 10.1145/1390156.1390290. 14

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018. 15

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 4911–4920. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00487. URL https://doi.org/10.1109/CVPR52688.2022.00487. 1, 2, 3, 5, 7, 8, 15, 17

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 681–688, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf. 4, 14

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 15

Xiao, Z., Kreis, K., Kautz, J., and Vahdat, A. VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models. In *Proceedings of International Conference on Learning Representations*, 2021. ISBN 2010.00654v2. URL http://arxiv.org/abs/2010.00654. 3

Xie, J., Lu, Y., Zhu, S., and Wu, Y. N. A theory of generative convnet. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2635–2644. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/xiec16.html. 2

Xie, J., Lu, Y., Gao, R., and Wu, Y. N. Cooperative learning of energy-based model and latent variable model via MCMC teaching. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 4292–4301, 2018. 3

Xie, J., Zheng, Z., and Li, P. Learning energy-based model with variational auto-encoder as amortized sampler. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 10441–10451. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17250. 3

Xie, J., Zheng, Z., Fang, X., Zhu, S., and Wu, Y. N. Cooperative training of fast thinking initializer and slow thinking solver for conditional learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):3957–3973, 2022a. doi: 10.1109/TPAMI.2021.3069023. URL https://doi.org/10.1109/TPAMI.2021.3069023. 3

Xie, J., Zhu, Y., Li, J., and Li, P. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=31d5RLCUuXC. 3

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., and Liu, Z. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, number NeurIPS, pp. 1–13, 2022. URL http://arxiv.org/abs/2210.07242. 8, 15

Yin, Y., Le Guen, V., Dona, J., Ayed, I., de Bézenac, E., Thome, N., and Gallinari, P. Augmenting physical models with deep networks for complex dynamics forecasting. In *Ninth International Conference on Learning Representations ICLR 2021*, 2021. 3

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 15

Zeng, A., Song, S., Lee, J., Rodriguez, A., and Funkhouser, T. A. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 36:1307–1319, 2020. 3

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 15

Zisselman, E. and Tamar, A. Deep Residual Flow for Out-of-Distribution Detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 13991–14000, 2020. doi: 10.1109/CVPR42600.2020.01401. 3, 5

A. HEAT Method

A.1. Energy-based models

An energy-based model (EBM) is an unnormalized density model defined via its energy function $E_\theta : \mathbb{R}^m \rightarrow \mathbb{R}$ parameterized by a neural network with parameters θ. For $\mathbf{z} \in \mathbb{R}^m$, its probability density is given by the Boltzmann distribution

$$p_\theta(\mathbf{z}) = \frac{1}{Z_\theta} \exp(\text{-}E_\theta(\mathbf{z})), \tag{8}$$

where Z_θ is the partition function which is intractable in high dimension. We can train EBMs via maximum likelihood estimation:

$$\underset{\theta}{\operatorname{argmax}} \log p_\theta(\mathcal{D}) = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\mathbf{z}\sim p_{in}}[-\log p_\theta(\mathbf{z})] \tag{9}$$

which can be approximated via stochastic gradient descent :

$$\theta_{i+1} = \theta_i - \lambda \nabla_\theta (\text{-}\log p_{\theta_i}(\mathbf{z})) \quad \text{with} \quad \mathbf{z} \sim p_{in} \tag{10}$$

Interestingly, $\nabla_\theta(\text{-}\log p_{\theta_i}(\mathbf{z}))$ can be computed without computing the intractable normalization constant Z_θ.

We have

$$\begin{aligned}
\nabla_\theta(\text{-}\log p_\theta(\mathbf{z})) &= \nabla_\theta E_\theta(\mathbf{z}) + \nabla_\theta \log Z_\theta \\
&= \nabla_\theta E_\theta(\mathbf{z}) + \frac{1}{Z_\theta} \nabla_\theta Z_\theta \\
&= \nabla_\theta E_\theta(\mathbf{z}) + \frac{1}{Z_\theta} \nabla_\theta \int_{\mathbf{z}} \exp(-E_\theta(\mathbf{z})) d\mathbf{z} \\
&= \nabla_\theta E_\theta(\mathbf{z}) + \frac{1}{Z_\theta} \int_{\mathbf{z}} \nabla_\theta \exp(-E_\theta(\mathbf{z})) d\mathbf{z} \\
&= \nabla_\theta E_\theta(\mathbf{z}) + \int_{\mathbf{z}} -\nabla_\theta E_\theta(\mathbf{z}) \frac{\exp(-E_\theta(\mathbf{z}))}{Z_\theta} d\mathbf{z} \\
&= \nabla_\theta E_\theta(\mathbf{z}) - \mathbb{E}_{\mathbf{z}'\sim p_\theta}[\nabla_\theta E_\theta(\mathbf{z}')].
\end{aligned}$$

Therefore, training EBMs via maximum likelihood estimation (MLE) amounts to perform stochastic gradient descent with the following loss:

$$\mathcal{L}_{MLE} = \mathbb{E}_{\mathbf{z}\sim p_{in}}\big[E_\theta(\mathbf{z})\big] - \mathbb{E}_{\mathbf{z}'\sim p_\theta}\big[E_\theta(\mathbf{z}')\big]. \tag{11}$$

Intuitively, this loss amounts to diminishing the energy for samples from the true data distribution $p(x)$ and to increasing the energy for synthesized examples sampled according from the current model. Eventually, the gradients of the energy function will be equivalent for samples from the model and the true data distribution and the loss term will be zero.

The expectation $\mathbb{E}_{\mathbf{z}'\sim p_\theta}\big[E_\theta(\mathbf{z}')\big]$ can be approximated through MCMC sampling, but we need to sample z' from the model p_θ which is an unknown moving density. To estimate the expectation under p_θ in the right hand-side of equation (11) we must sample according to the energy-based model p_θ. To generate synthesized examples from p_θ, we can use gradient-based MCMC sampling such as Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011) or Hamiltonian Monte Carlo (HMC) (Neal, 2010). In this work, we use SGLD sampling following (Du & Mordatch, 2019; Grathwohl et al., 2020). In SGLD, initial features are sampled from a proposal distribution p_0 and are updated for T steps with the following iterative rule:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \frac{\eta}{2} \nabla_{\mathbf{z}} E^h_{\theta_k}(\mathbf{z}_t) + \sqrt{\eta}\, w_t, \text{ with } w_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{12}$$

where η is the step size. Therefore sampling from p_θ does not require to compute the normalization constant Z_θ either.

Many variants of this training procedure have been proposed including Contrastive Divergence (CD) (Hinton, 2002) where $p_0 = p_{\text{data}}$, or Persistent Contrastive Divergence (PCD) (Tieleman, 2008) which uses a buffer to extend the length of the MCMC chains. We refer the reader to (Song & Kingma, 2021) for more details on EBM training with MLE as well as other alternative training strategies (score-matching, noise contrastive estimation, Stein discrepancy minimization, etc.).

A.2. Composition function

While many composition strategies can be consider, we choose to use the best trade-off between detection efficiency and flexibility. While combining many OOD-prior is great, hand tuning many balancing hyper-parameters can quickly become cumbersome. Our energy composition strategy E_{HEAT}^{β} presents the advantage to have only one hyper-parameter β to tune with a clear interpretation for its different regimes. Indeed, depending on β, this composition operator generalizes several standard aggregation operators. When $\beta \to +\infty$, we recover the maximum operator, while when $\beta \to -\infty$, we recover the minimum operator. In the $\beta \to 0$ case, we recover the sum and the resulting distribution is equivalent to a product of experts. Finally, taking $\beta = -1$ amounts to using the *logsumexp* operator which approximates a mixture of experts. In addition, to prevent the energy of one prior scorer to dominate the others, we normalize the energies using the train statistics (subtracting their mean and dividing by their standard deviation). This simple standardization gives good results in our setting, although more advanced normalization schemes could certainly be explored if needed with other prior scorers.

B. Experiments

Datasets. We conduct experiments using CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009) as in-distribution datasets. For OOD datasets, we define three categories: near-OOD datasets, mid-OOD datasets and far-OOD datasets. These correspond to different levels of proximity with the ID datasets. For CIFAR-10 (resp. CIFAR-100), we consider `TinyImagenet`[3] and CIFAR-100 (resp. CIFAR-10) as near-OOD datasets. Then for both CIFAR-10 and CIFAR-100, we use `LSUN` (Yu et al., 2015) and `Places` (Zhou et al., 2017) datasets as mid-OOD datasets, and `Textures` (Cimpoi et al., 2014) and `SVHN` (Netzer et al., 2011) as far-OOD datasets. We use different Imagenet (Deng et al., 2009) benchmarks. In Tab. 5 we use the benchmarks of (Sehwag et al., 2021; Sun et al., 2022) with `iNaturalist` (Van Horn et al., 2018), `LSUN` (Yu et al., 2015), `Places` (Zhou et al., 2017) and `Textures` (Cimpoi et al., 2014). In Appendix B.1.1 we use the Imagent benchmark recently introduced in OpenOOD (Yang et al., 2022), and we refer the reader to the paper for details about the dataset[4]. Finally in Appendix B.1.2 we ue the Imagenet benchmark proposed in (Wang et al., 2022), with notably the OpenImage-O dataset introduced specifically as an OOD dataset for the Imagenet benchmark in (Wang et al., 2022), for more details we refer the reader to the paper.

Implementation details. All experiments were conducted using `PyTorch` (Paszke et al., 2019). We use a ResNet-34 classifier from the `timm` library (Wightman, 2019) for the CIFAR-10 and CIFAR-100 datasets and a ResNet-50 for the Imagenet experiments. HEAT consists in a 6 layers MLP trained for 20 epochs with Adam with learning rate $5e$-6. The network input dimension is 512 (which is the dimension of the penultimate layer of ResNet-34) for the CIFAR-10/100 benchmarks and 2048 (which is the dimension of the penultimate layer of ResNet-50) for the Imagenet benchmark. The hidden dimension is 1024 for CIFAR-10/100 and 2048 for Imagenet, and the output dimension is 1. For SGLD sampling, we use 20 steps with an initial step size of $1e$-4 linearly decayed to $1e$-5 and an initial noise scale of $5e$-3 linearly decayed to $5e$-4. We add a small Gaussian noise with std $1e$-4 to each input of the EBM network to stabilize training as done in previous works (Du & Mordatch, 2019; Grathwohl et al., 2020). The L_2 coefficient is set to 10. We use temperature scaling on the mixture of Gaussian distributions energy with temperature $T_{\mathcal{G}} = 1e3$. The hyper-parameters for the CIFAR-10 and CIFAR-100 models are identical.

Additional metric In Appendix B.1.1 we use an additional metric the AUPR, which measures the area under the Precision-Recall (PR) curve, using the ID samples as positives (see (Yang et al., 2022) for details). The score corresponds to the AUPR-In metric in other works.

B.1. Additional comparison to state-of-the-art

B.1.1. IMAGENET OPENOOD RESULTS

We compare HEAT on the recently introduced OpenOOD benchmark (Yang et al., 2022) in Tabs. 6 to 8. In addition to the baselines used in the main paper, the OpenOOD benchmark includes the Mahalanobis detector (Lee et al., 2018b) (MDS), OpenMax (Bendale & Boult, 2016), the Gram matrix detector (Sastry & Oore, 2020) (Gram), KL matching (Hendrycks et al., 2022) (KLM) and GradNorm (Huang et al., 2021). We show that on the aggregated results in Tab. 6 HEAT outperforms previous methods, and sets new state-of-the-art performances for our considered setting. Similarly to our comparison in Sec. 4.2 we can see that HEAT performs well on far-OOD Tab. 7 and near-OOD Tab. 8. On far-OOD HEAT has the best performances on each metric, *i.e.* 2.4 FPR95, 99.4 AUC and 100 AUPR. On near-OOD HEAT has the best performances on

[3] The dataset can be found at: `https://www.kaggle.com/c/tiny-imagenet`

[4] Datasets for the OpenOOD benchmark can be downloaded using: `https://github.com/Jingkang50/OpenOOD`.

AUC, *i.e.* +2.6 pts AUC *vs.* KNN, -0.8 pts FPR95 and +0.9 pts AUPR *vs.* ReAct.

Table 6. **Aggregated results on the Imagenet OpenOOD benchmark**. All methods are based on an Imagnet pre-trained ResNet-50.

Method	Near-OOD	Far-OOD	Average
	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑
OpenMax	76.3 / 66.0 / 92.2	55.0 / 84.9 / 97.0	69.2 / 72.3 / 93.8
MSP	73.7 / 69.3 / 94.6	57.8 / 86.2 / 97.5	68.4 / 74.9 / 95.6
ODIN	68.2 / 73.2 / 95.0	21.8 / 94.4 / 99.1	52.7 / 80.2 / 96.3
MDS	86.9 / 68.3 / 90.8	18.4 / 94.0 / 98.5	64.1 / 76.8 / 93.3
Gram	83.1 / 68.3 / 91.8	43.2 / 89.2 / 97.9	69.8 / 75.3 / 93.8
EL	73.3 / 73.5 / 95.3	33.8 / 92.8 / 98.8	60.1 / 79.9 / 96.4
GradNorm	61.1 / 75.7 / 94.9	15.0 / 95.8 / 99.3	45.7 / 82.4 / 96.3
ReAct	57.2 / 79.3 / 96.2	23.8 / 95.2 / 99.3	46.1 / 84.6 / 97.2
MLS	72.2 / 73.6 / 95.3	37.6 / 92.3 / 98.7	60.7 / 79.8 / 96.5
KLM	68.1 / 73.4 / 94.8	56.6 / 88.8 / 98.1	64.3 / 78.5 / 95.9
VIM	73.8 / 79.9 / 95.7	6.9 / 98.4 / **99.8**	51.5 / 86.1 / 97.1
KNN	71.9 / 80.8 / 95.7	8.4 / 98.0 / **99.7**	50.8 / 86.5 / 97.0
DICE	65.1 / 73.8 / 95.1	15.8 / 95.7 / 99.3	48.6 / 81.1 / 96.5
HEAT	**55.2 / 84.8 / 97.1**	**2.6 / 99.4 / 100.0**	**37.7 / 89.6 / 98.1**

Table 7. **Results on far-OOD of the Imagenet OpenOOD benchmark**. All methods are based on an Imagnet pre-trained ResNet-50.

Method	**Textures**	**MNIST**	**Far-OOD**
	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑
OpenMax	65.3 / 78.9 / 96.0	44.6 / 90.9 / 98.0	55.0 / 84.9 / 97.0
MSP	63.6 / 82.5 / 97.2	52.0 / 89.8 / 97.8	57.8 / 86.2 / 97.5
ODIN	42.5 / 89.2 / 98.3	**0.9 / 99.7 / 99.9**	21.8 / 94.4 / 99.1
MDS	36.7 / 90.2 / 97.4	**0.0** / 97.7 / **99.6**	18.4 / 94.0 / 98.5
Gram	53.3 / 82.7 / 96.7	33.1 / 95.7 / 99.2	43.2 / 89.2 / 97.9
EL	49.2 / 88.8 / 98.3	18.3 / 96.8 / 99.4	33.8 / 92.8 / 98.8
GradNorm	29.4 / 92.1 / 98.7	**0.6 / 99.5 / 99.9**	15.0 / 95.8 / 99.3
ReAct	43.1 / 91.6 / 98.9	4.5 / 98.8 / 99.8	23.8 / 95.2 / 99.3
MLS	51.3 / 88.5 / 98.3	24.0 / 96.1 / 99.2	37.6 / 92.3 / 98.7
KLM	67.6 / 84.7 / 97.6	45.5 / 92.9 / 98.6	56.6 / 88.8 / 98.1
VIM	12.4 / 97.5 / **99.7**	1.4 / 99.2 / **99.9**	6.9 / 98.4 / **99.8**
KNN	16.9 / 96.2 / **99.5**	**0.0 / 99.9 / 99.9**	8.4 / 98.0 / **99.7**
DICE	29.4 / 92.1 / 98.7	2.3 / 99.3 / **99.9**	15.8 / 95.7 / 99.3
HEAT	**5.3 / 98.9 / 99.9**	**0.0 / 99.9 / 100.0**	**2.6 / 99.4 / 100.0**

Table 8. **Results on near-OOD of the Imagenet OpenOOD benchmark**. All methods are based on an Imagnet pre-trained ResNet-50.

Method	**Species**	**iNaturalist**	**OpenImage-O**	**Imagenet-O**	**Near-OOD**
	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑	FPR95↓ / AUC ↑ / AUPR ↑
OpenMax	81.8 / 70.8 / 90.5	56.8 / 79.5 / 92.7	66.7 / 81.5 / 91.4	99.9 / 32.1 / 94.1	76.3 / 66.0 / 92.2
MSP	79.2 / 75.2 / 92.9	52.4 / 88.5 / 97.1	63.2 / 85.0 / 94.2	100.0 / 28.7 / 94.3	73.7 / 69.3 / 94.6
ODIN	80.5 / 71.6 / 91.4	42.0 / 91.2 / 97.7	50.5 / 88.4 / 95.3	99.9 / 41.5 / 95.4	68.2 / 73.2 / 95.0
MDS	94.8 / 60.2 / 87.6	93.7 / 67.8 / 90.9	82.4 / 70.7 / 86.4	**76.9 / 74.3 / 98.2**	86.9 / 68.3 / 90.8
Gram	86.8 / 67.3 / 89.9	75.2 / 78.4 / 94.1	79.0 / 71.9 / 86.8	91.5 / 55.8 / 96.4	83.1 / 68.3 / 91.8
EL	82.3 / 72.1 / 91.6	53.7 / 90.6 / 97.8	57.0 / 89.2 / 96.0	100.0 / 42.0 / 95.7	73.3 / 73.5 / 95.3
GradNorm	74.4 / 75.7 / 92.8	26.9 / 93.9 / 98.4	47.5 / 85.2 / 92.8	95.6 / 48.2 / 95.5	61.1 / 75.7 / 94.9
ReAct	**68.4 / 77.5 / 92.6**	**19.3 / 96.4 / 99.2**	43.5 / 90.6 / 96.4	98.1 / 52.5 / 96.6	57.2 / 79.3 / 96.2
MLS	80.8 / 73.0 / 91.8	50.8 / 91.2 / 97.9	57.1 / 89.3 / 96.0	100.0 / 41.0 / 95.6	72.2 / 73.6 / 95.3
KLM	**73.7** / 74.5 / 91.7	41.1 / 90.8 / 97.4	57.8 / 87.4 / 94.7	100.0 / 40.8 / 95.4	68.1 / 73.4 / 94.8
VIM	84.0 / 70.7 / 90.9	68.0 / 88.4 / 97.4	57.7 / 89.6 / 96.4	85.5 / 70.9 / **98.3**	73.8 / 79.9 / 95.7
KNN	76.4 / 76.4 / **93.0**	68.6 / 85.0 / 96.4	58.0 / 86.4 / 94.9	84.8 / 75.4 / **98.6**	71.9 / 80.8 / 95.7
DICE	78.6 / 71.3 / 91.3	35.3 / 92.5 / 98.1	47.7 / 88.5 / 95.3	98.5 / 42.9 / 95.5	65.1 / 73.8 / 95.1
HEAT	74.3 / 76.8 / **93.7**	27.4 / 95.0 / 98.9	**41.4 / 91.8 / 97.2**	77.6 / 75.5 / **98.7**	**55.2 / 84.8 / 97.1**

B.1.2. ViT results

In Tab. 9 we compare HEAT using a Vision Transformer[5] (ViT), on the Imagenet benchmark introduced in (Wang et al., 2022). We show that on the aggregated results HEAT outperforms the previous best method, ViM (Wang et al., 2022), by -1.7 pts FPR95. Importantly HEAT ouperforms other method on three datasets of the benchmark, *i.e.* OpenImage-O, Textures, Imagenet-O, and is competitive on iNaturalist. Tab. 9 demonstrates the ability of HEAT to adapt to architectures of neural networks, *i.e.* Vision Transformer (Dosovitskiy et al., 2020), other than the convolutional networks (*i.e.* ResNet-34 & ResNet-50) tested in Sec. 4.2.

Table 9. **Results on Imagenet.** All methods are based on an Imagenet pre-trained **Vision Transformer** (ViT) model. ↑ indicates larger is better and ↓ the opposite.

Method	**OpenImage-O**	**Textures**	**iNaturalist**	**Imagenet-O**	**Average**
	FPR95↓ / AUC↑	FPR95↓ / AUC↑	FPR95↓ / AUC↑	FPR95↓ / AUC↑	FPR95↓ / AUC↑
MSP	34.2 / 92.5	48.6 / 87.1	19.0 / 96.1	64.8 / 81.9	41.7 / 89.4
EL	14.0 / 97.1	28.2 / 93.4	6.2 / 98.7	41.3 / 90.5	22.4 / 94.9
ODIN	15.7 / 96.9	30.6 / 93.0	6.6 / 98.6	44.2 / 89.9	24.3 / 94.6
MaxLogit	15.7 / 96.9	30.6 / 93.0	6.6 / 98.6	44.2 / 89.9	24.3 / 94.6
KL Matching	28.5 / 93.9	44.1 / 88.8	14.8 / 96.9	55.7 / 84.1	35.8 / 90.9
KNN	45.8 / 91.7	28.9 / 93.2	52.3 / 91.1	52.9 / 88.4	45.0 / 91.1
Residual	32.6 / 92.7	33.8 / 92.2	6.6 / 98.6	47.9 / 88.2	30.2 / 92.9
ReAct	13.5 / 97.4	28.5 / 93.3	4.3 / 99.0	42.6 / 90.7	22.2 / 95.1
Mahalanobis	13.5 / 97.5	25.2 / 94.2	**2.1 / 99.5**	37.0 / 92.8	19.5 / 96.0
ViM	12.6 / 97.6	20.3 / 95.3	2.6 / **99.4**	36.8 / 92.6	18.1 / 96.2
HEAT	**11.2 / 97.8**	**12.8 / 96.9**	6.9 / 98.2	**34.8 / 93.1**	**16.4 / 96.5**

B.1.3. Results Imagenet with DICE

In this section we compare the performance of HEAT when using DICE instead of EL as one its components. We show in Tab. 10 that using DICE instead of EL as part of HEAT's components further improves the OOD detection performances on all OOD datasets except Textures.

Table 10. **Results of HEAT using DICE vs EL on Imagenet.** All methods use an Imagenet pre-trained ResNet-50. Results are reported with FPR95↓ / AUC↑.

Method	**iNaturalist**	**SUN**	**Places**	**Textures**	**Average**
HEAT (w/. EL)	28.1 / 94.9	44.6 / 90.7	58.8 / 86.3	**5.9 / 98.7**	34.4 / 92.6
HEAT (w/. DICE)	**24.4 / 95.4**	**39.5 / 91.4**	**53.2 / 87.4**	9.7 / 97.5	**31.7 / 93.1**

B.1.4. Results Supervised Contrastive backbone

In this section we evaluate HEAT when using a supervised contrastive backbone and compare with KNN+ and SSD+. We show in Tab. 11 that HEAT still largely outperforms the competition even with the supervised contrastive backbone.

Table 11. **Results on Imagenet.** All methods use an Imagenet Supervised Contrastive pre-trained ResNet-50. Results are reported with FPR95↓ / AUC↑.

Method	**iNaturalist**	**SUN**	**Places**	**Textures**	**Average**
SSD+	34.3 / 95.0	65.2 / 86.3	70.2 / 83.8	14.7 / 95.6	46.1 / 90.2
KNN+	30.8 / 94.7	48.9 / 88.4	60.0 / 84.6	17.0 / 94.5	39.2 / 90.6
HEAT	**14.6 / 97.2**	**32.4 / 93.4**	**45.4 / 89.9**	**9.7 / 97.7**	**25.5 / 94.6**

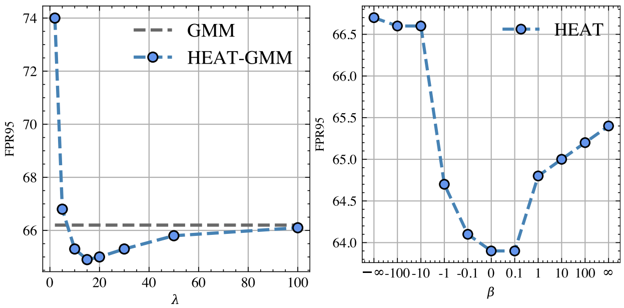
[5] The model used can be found at https://github.com/haoqiwang/vim

B.2. Model analysis

In Fig. 6 we show the impact of λ in Eq. (6) and β *vs.*FPR95 on CIFAR-100, we study in Fig. 7 how HEAT behaves on low data regimes with CIFAR-10 as ID dataset. Finally in Tab. 12 we study the computational requirements of HEAT.

Robustness to λ In Fig. 6a we can see that we have similar trends to Fig. 3a. For values of λ too high, *i.e.* when the expressivity of the energy-based correction is limited, HEAT-GMM has the same performances than GMM. For values of λ too low the energy-based correction is not controlled and disregards the prior scorer, *i.e.*GMM. Finally for a wide range of λ values HEAT-GMM improves the OOD detection performances of GMM.

Robustness to β In Fig. 6b we show that HEAT is stable wrt. β on CIFAR-100 similarly to Fig. 3b.



(a) λ *vs.*FPR95↓

(b) β *vs.*FPR95↓

Figure 6. On CIFAR-100 ID: (a) impact of λ in Eq. (6) *vs.* FPR95 and (b) analysis of β in Eq. (7) *vs.* FPR95.

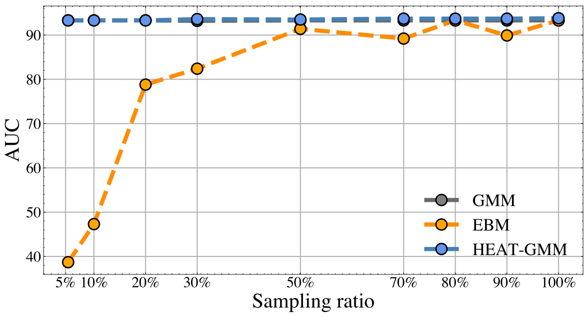


Figure 7. Impact on performances (AUC↑ on CIFAR-10) *vs.* the number of training data for GMM density, fully data-driven EBM, and HEAT. Our hybrid approach maintains strong performances in low-data regime, in contrast to the fully data-driven EBM.

Low data regime Similarly to Fig. 4, we can see that training solely an EBM is very unstable when the number of data is low. On the other hand HEAT-GMM is stable to the lack of data and improves GMM even with few ID samples available.

Computational cost In Tab. 12 we report the cost of computing of different components of HEAT, *e.g.* forward pass of a ResNet-50, energy computation of GMM. We extrapolate based on those inference time the computational cost of deep-ensembles (Lakshminarayanan et al., 2017) and of HEAT. The compute time for HEAT, 5.194ms, is due at 84% by the inference time of a ResNet-50. This is why deep-ensembles has a compute time of 2500ms which is 4.8 times larger than that of HEAT. Further more we can see that correcting GMM with HEAT only brings an overhead of 1ms, which will not scale with the size of the model but only its embedding size, *e.g.* CLIP (Radford et al., 2021) as an embedding size of 1024 for its largest model.

Table 12. Computational cost reported in **ms** ↓. Times are reported using RGB images of size 224 × 224, a ResNet-50 with an output size of 2048, 1000 classes (*i.e.* Imagenet setup), on a single GPU (Quadro RTX 6000 with 24576MiB).

ResNet-50	GMM	GMM-std	EL	EBM	deep-ensemble	**HEAT**
500	8	8	0.4	1	2501	519.4

B.3. Qualitative results

We show qualitative results of HEAT *vs.* EL (Liu et al., 2020) and SSD (Sehwag et al., 2021) on LSUN Fig. 8 and Textures Fig. 9. On Fig. 8 we can see that EL and SSD detect different OOD samples. HEAT is able though the correction and composition to recover those mis-detected OOD samples. On Fig. 9 we can see that SSD performs well on the far-OOD dataset (Textures), however HEAT is able to recover a mis-detected OOD sample. Fig. 8 and Fig. 9 qualitatively show how HEAT is able to better mis-detect OOD samples.

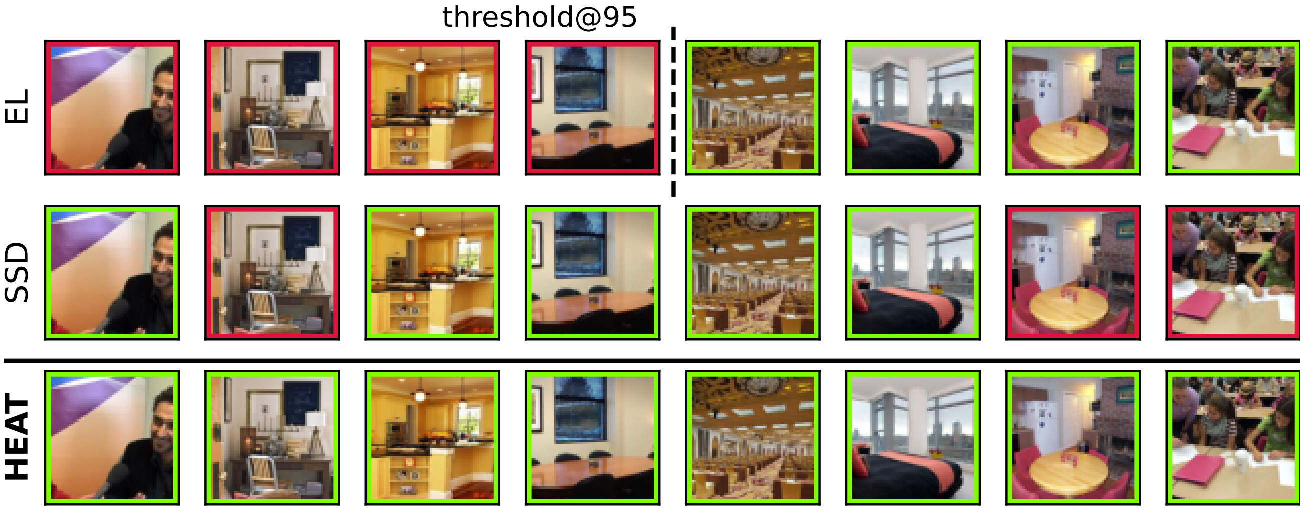


Figure 8. Qualitative comparison of HEAT *vs.* EL (Liu et al., 2020) and SSD (Sehwag et al., 2021) on **LSUN**. Samples in green are correctly detected as OOD (above the 95% of ID threshold), samples in red are incorrectly predicted as ID, *i.e.* an energy lower than the threshold.

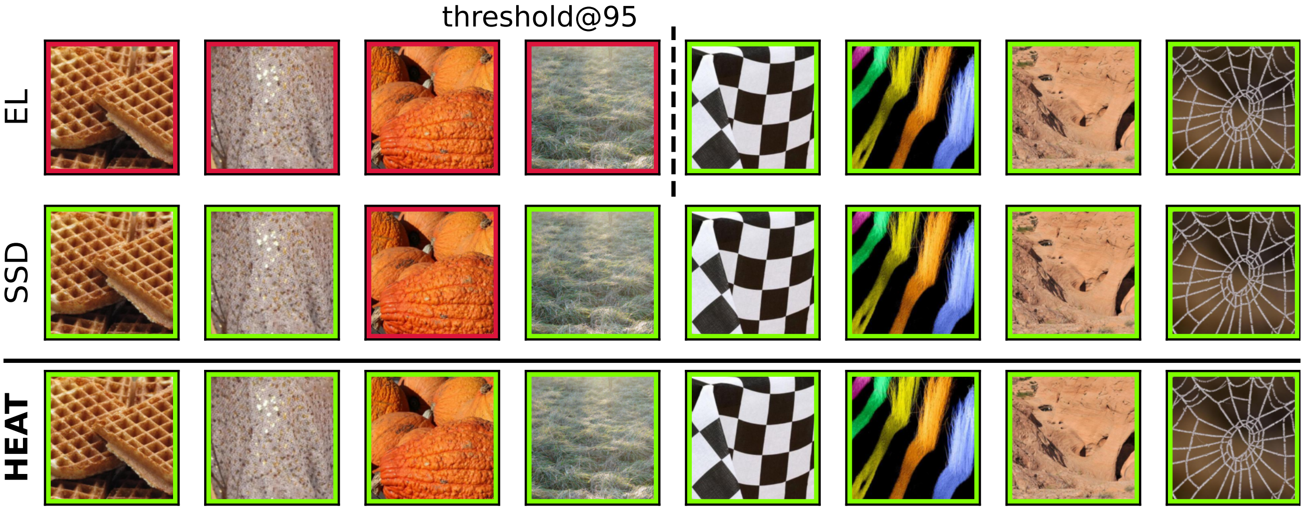


Figure 9. Qualitative comparison of HEAT *vs.* EL (Liu et al., 2020) and SSD (Sehwag et al., 2021) on **Textures**. Samples in green are correctly detected as OOD (above the 95% of ID threshold), samples in red are incorrectly predicted as ID, *i.e.* an energy lower than the threshold.