SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Games Engineering

Out-of-Distribution Detection via Post-hoc Decoupling Semantic Segmentation

Barış Zöngür

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Games Engineering

Out-of-Distribution Detection via Post-hoc Decoupling Semantic Segmentation

Out-of-Distribution-Erkennung über Post-hoc-Entkopplung der Semantischen Segmentierung

Author:	Barış Zöngür
Supervisor:	Prof. Dr. Daniel Cremers
Advisor:	Dr. Nikita Araslanov, Simon Weber
Submission Date:	15/05/2024

I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, 15/05/2024 Barış Zöngür

Acknowledgments

First and foremost, I would like to thank Dr. Nikita Araslanov and Simon Weber for their close supervision and guidance throughout my thesis and my employment as a research assistant.

Secondly, I would like to thank Prof. Dr. Daniel Cremers for providing me with the opportunity to write my thesis in the refined and resourceful environment of Computer Vision Group(CVG).

Lastly, I would like to thank my parents for their unwavering emotional and financial support throughout my studies.

Abstract

Real-world deployment of segmentation models elevates the objective of out-of-distribution (OOD) detection. Models that utilize OOD supervision implicitly extend the in-distribution (ID) set, contradicting the fundamentals of the objective. On the other hand, unsupervised methods rely on inferring the inherent information from the model's decision patterns. Recent methods utilize class scores as an informative medium for OOD detection. Surprisingly, the role that the embedding positions play on the class scores is mainly unexplored. Our technical analysis and empirical observations uncover that the norm of the embeddings contains information that is primarily related to the distribution of the external factors, regardless of the objects' whatness. Consequently, norms aggravate ambiguity in the class scores, rendering them less informative as a comparative measure. To mitigate the external ambiguity, we propose normalized class scores obtained by post-hoc projection of embeddings into a hypersphere. Additionally, our observations suggest that class weight normalization utilized in the classification does not extend to the segmentation domain. Through empirical evaluation, we show that unsupervised OOD detection methods employed with normalized class scores consistently outperform their counterparts for multiple datasets and different backbones. Furthermore, we identify the driving factor of the ViT model's OOD accuracy as their norm uniformity by showing the practical equivalence of ViT models to hyperspherical representation.

Contents

1 Introduction

Advancements in semantic segmentation models enabled a medium where such models can be used in real-world scenarios [44, 34, 11, 10]. With extended usage, it is now apparent that further advancement is not only bottlenecked by the discriminative power of the model but also the assumptions. By constraints of segmentation models, the correctness of a prediction can only be defined if prior knowledge of the testing domain is assumed. However, with the widened usage of segmentation models, the assumption of equivalence of the In-Distribution (ID) set and the global set of all possible use cases can not hold anymore. By defining the set difference between global and ID sets as Out-Of-Distribution (OOD), OOD detection models aim to discriminate between ID and OOD sets [27, 24, 32, 7, 29, 33, 37].

In recent literature, models commonly try to approximate the OOD set as a union of numerous predefined low-level classes [7, 48, 2, 18, 33]. These non-intersecting low-level classes are manually selected from a distinct classification taxonomy and used to supervise the model to learn this union set as OOD. As a result, instead of discriminating between ID and OOD sets, such models learn to classify predefined subsets of the OOD set. This selection is equivalent to selecting a subset from the global set and classifying the selected subset using direct supervision. In segmentation tasks, ID objects are defined as a selected subset of the global set, which is then learned by providing supervision. One can argue that such OOD-supervised models extend the ID set by adding a subset of the global set rather than estimating the difference between the global set and the ID set.

Conversely, unsupervised models use inherent information from the model's decision to estimate a score function for OOD detection [32, 25, 28, 23]. In deep models, this information is generally extracted from the last layer of the model, where the representation of the input is converged. In supervised and unsupervised models alike, class scores or logits are commonly interpreted as a representative medium

to detect if an input belongs to any ID distributions [32, 18, 48, 33, 51, 23, 25, 56, 38, 37]. For the same input, the internal ranking of the class scores determines the prediction confidence. Consequently, a high class score compared to other classes yields high confidence. However, this does not indicate that class scores are an informative medium for comparing different inputs. A low class score can produce high confidence if its exponential ranking is superior to that of other classes. Inversely, high class scores can result in low confidence if class scores are uniform.

In this work, our contributions are as follows. *(i)* We challenge the class score based unsupervised OOD detection methods through technical investigation of components of class score generation. *(ii)* We identify the norm as a degree of freedom the model uses to stay robust under different external shifts. *(iii)* Decoupled from the object's whatness, we claim the norm as the driving factor of the ambiguity in class scores. *(iv)* Via post-hoc mapping of embedding to a hypersphere, we mitigate the ambiguity imposed by the norms. *(v)* We reason the high OOD detection accuracy of ViT models to their practical equivalence to a hypersherical representation. *(vi)* With thorough evaluation, we support our claims with empirical evidence.

2 Preliminaries

2.1 Semantic Segmentation

Semantic segmentation is the problem of clustering pixels in a given image where a semantic taxonomy defines each cluster [39]. Pixels corresponding to the same class of objects clustered together to achieve a human-understandable partitioning in the image. If an expected taxonomy over objects is unknown, clusters obtained by segmentation are used to infer the information about the possible semantic classes in a given domain by leveraging found object similarities. Such methods are called unsupervised segmentation methods [55, 21, 13]. However, more commonly, an expected taxonomy on the domain is known, and either the clusters are found to match the predefined classes after processing, or clusters are predefined by matching semantic classes before processing. The decision to use pre- or post-process matching generally depends on the availability of labeled data for a subset of the domain [36]. Methods that leverage the labeled data to define semantic classes and supervise the representation of pixels to match the predefined class distribution are called supervised segmentation methods [47, 54, 9, 12, 50].

The supervised classification aims to extract a singular representation for each data point, assuming the data points are independent [3]. In the segmentation case, the class of each pixel also depends on the neighboring pixels and global information about the image [8, 47]. Such dependence raises the need for a segmentation model to generate a representation for each model containing the neighboring information. Aligned with the objective, convolutional models [45, 22] enabled a medium for the segmentation task, extending their usage to real-world scenarios. Later, with the discovery of the applicability of attention models [49] in the visual domain and, as a result, representations that are further supervised by global information, the reach of the segmentation models was extended [5, 40, 44, 34].

Even though segmentation differs in many aspects from the task of classification, it contains the same closed-world assumption. For a given pixel x, segmentation aims to estimate the probabilities $p(x \in c_i|x)$, or in short notation $p(c_i|x)$, where $C = \{c_1, c_2, ..c_N\}$ is the set of known classes with $|C| = N$. The closed set assumptions follow that:

$$\sum_{i}^{N} p(c_i|x) = 1 \tag{2.1}$$

Note that this assumption directly implies $G \setminus \bigcup_i^N c_i = \emptyset$ where G denotes the global set. One can calculate the probabilities $p(c_i|x)$ if $p(x|c_i)$ is known as Bayesian formula follows:

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{\sum_j^N p(x|c_j)p(c_j)} \tag{2.2}$$

Here, another assumption over class cardinalities are made as $|c_i| = |c_j|, \forall c_i, c_j \in C$, which implies $p(c_i) = \frac{1}{N}, \forall c_i \in C$. Following the assumption, we can rewrite eq 2.2 as:

$$p(c_i|x) = \frac{p(x|c_i)}{\sum_j^N p(x|c_j)} \tag{2.3}$$

In a deep model, the log of $p(x|c_i)$ is estimated by a learned function $f(w, x)$ where w corresponds to the model parameters. With the estimation, probabilities recovered by:

$$p(c_i|x) = \frac{e^{f(w_i,x)}}{\sum_j^N e^{f(w_j,x)}} \tag{2.4}$$

Eq. 2.4 is referred as the softmax function in the literature where $f(w_i, x)$ are class scores or logits.

As a consequence of the assumptions of discriminative training, when such a model is moved to open-world settings, it only provides the predictions for predefined classes. Further, it assigns any objects to one of the constrained closed-world classes, which does not consistently yield a reasonable assignment in open-world settings where the assumed global set differs.

2.2 OOD Detection

The objective of OOD detection arises when the expected input distribution is shifted over the constrained class taxonomy [41]. With such a shift, the closed-world assumption on the input does not hold, and objects from outside the defined class taxonomy are expected. The assumption on the global set shifts as:

$$G = I \cup O \tag{2.5}$$

where $I \cap O = \emptyset$. Here O is the OOD set, and I is the ID set where $I = \bigcup_i^N c_i$. Considering the new assumptions about the input space, the task of OOD detection for a given input pixel x, is to find a measure for the OODness of the object that would parallel an estimation for $p(O|x)$. Said measure is generally a scalar score function $S(x)$ having a higher value for OOD inputs [32, 48, 33].

A common approach for OOD detection is providing OOD supervision for the model [48, 7, 16]. The assumed closed-world taxonomy for the objects defines the set ID set. The definition of OOD is the set difference between the global set and the ID set. Following the definition, models with OOD supervision sample points outside the ID distribution to estimate a sub-space for the OOD in the embedding space [33, 2]. Such models assume $C + 1$ classes in the input domain by extending the input set with a OOD "class". These models typically optimize a score function that would yield high values for OOD and low values for ID with additional training [32, 33, 18, 7, 48].

Our main argument against the supervised OOD methods is directly related to the very definition of an OOD. Defining the ID and OOD by the human expectations for the domain overlooks the decision mechanisms of supervised models. The distinction between ID and OOD with respect to a supervised model comes from the data that the model is supervised with. When OOD supervision is provided to a model, the distribution from which said OD samples are drawn becomes an ID distribution.

Other OOD detection methods do not use OOD supervision and aim to detect OOD inputs by inferring the inherent information from the model's decisions [23, 28, 24, 25, 1, 29]. Unsupervised methods differ by their level of invasiveness and generalizability. Invasive methods alter the model's training process to provide self-supervision. In the literature, such models are also called re-training methods. One common line

of invasive approach is generative methods, where OOD supervision is given by generated objects or noise [19, 31, 30]. Even though these methods can achieve high OOD detection scores, they also affect the ID accuracy of the model. Generalizability depends on factors such as the requirement of an extra model on top of the original model or the dependence of the method on a specific model. These methods offer a solution for when a specific type of model is used, thus reducing their impact on the extent of the method [28, 18, 37, 1]. One significant example is the recent trend of methods that are specific to the Mask2Former [11] model. Even though they can empirically show good results on the ood detection [18, 28], the proposed methods can not be used in any other segmentation method. Also, considering the interchangeability of methods between domains in the overall literature, such as their usage with other tasks or other types of data, being highly dependent on a specific model hinders their generalizability.

2.3 Backbone Architectures

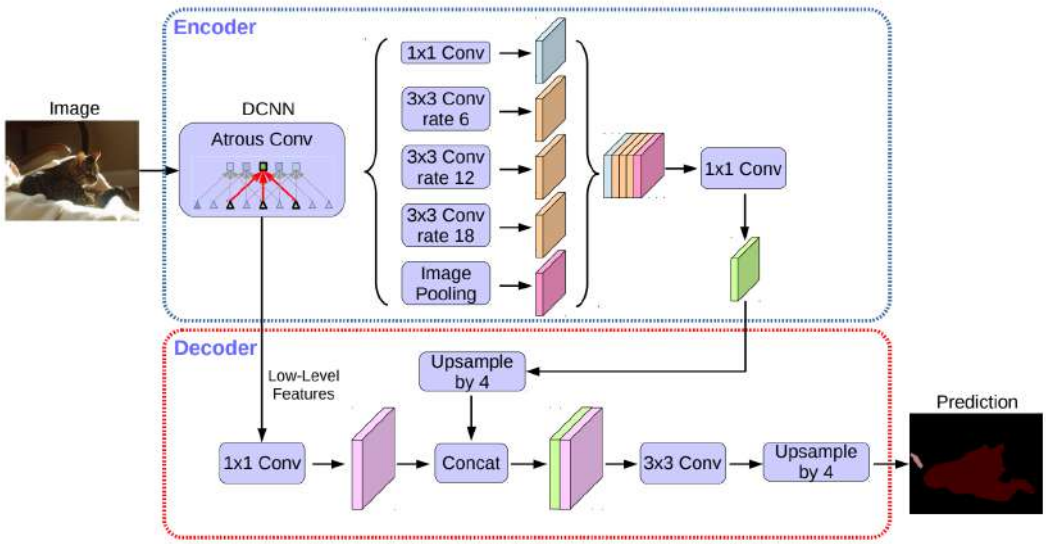

Figure 2.1: **Visualization of DeepLabv3+ architecture.** Image taken from the original source [10].

This part briefly overviews the different network architectures we employ during

OOD detection.

Figure 2.1 shows the architecture of **DeepLabv3+** [10], which enhances DeepLabv3 [9] with a decoder that utilizes the high field-of-view representations extracted from the DeepLabv3 backbone. To extract a representation for each pixel, DeepLabv3 uses atrous [52] convolutions, in which an atrous rate increases the stride of sampling in the convolution kernel. Further extending this, DeepLabv3+ employ ASPP(Atrous Spatial Pyramid Pooling), in which different representations with increasing atrous rates are concatenated. DeepLabv3+ upsamples the extracted features from the DeepLabv3 backbone with bilinear interpolation and combines them with low-level features. We use DeepLabv3+ architecture with ResNet-101 [22] backbone.

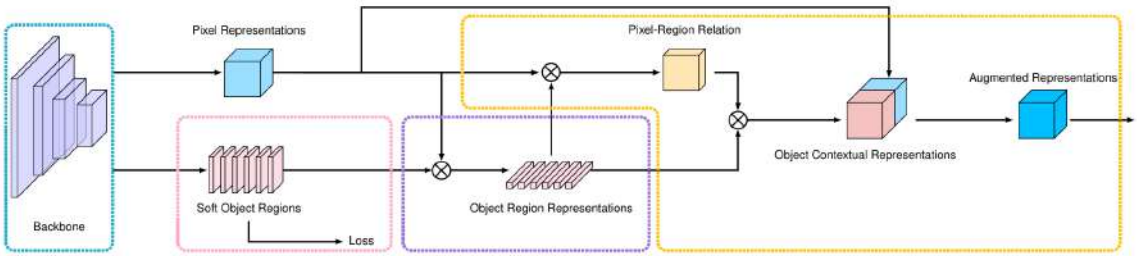

Figure 2.2: **Visualization of OCRNet architecture.** Image taken from the original source [54].

OCRNet [54] enhances the feature representations of the pixels with context-aware representations of each class region on the image. Figure 2.2 illustrates the architecture of OCRNet. They estimate soft object regions in the image where each region corresponds to a class. For each object region, they compute an object region representation by aggregating the features of each pixel belonging to the region. They extract the context-aware scores for each pixel by utilizing the similarity score between each pixel feature and their corresponding region score. They concatenate the enhanced features with the initial features to generate the final representation. Similarly, we use the OCRNet model with the ResNet-101 [22] backbone.

Mask2Former [11] is a segmentation model that can perform instance, panoptic, and semantic segmentation. It is built on the Maskformer [12] architecture, which outputs a binary mask for each class and optimizes an objective function that includes both classification and binary mask loss. They also utilize a Transformer decoder, aggregating global information from the latent space to all masks. Mask2Former, in

Figure 2.3, enhances the Transformer decoder by feeding representations from the decoder at multiple scales. They also introduce masked attention to extract localized features within predicted mask regions. Masked attention is similar to object regions in OCRNet, where attention is applied to the regions based on their semantics.

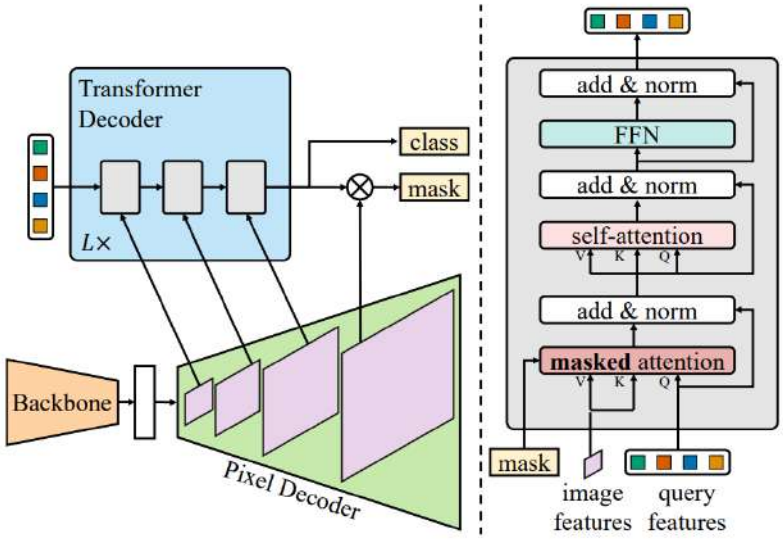

Figure 2.3: **Illustration of the Mask2Former architecture.** Image taken from the original source [11].

2.4 Cityscapes Taxonomy

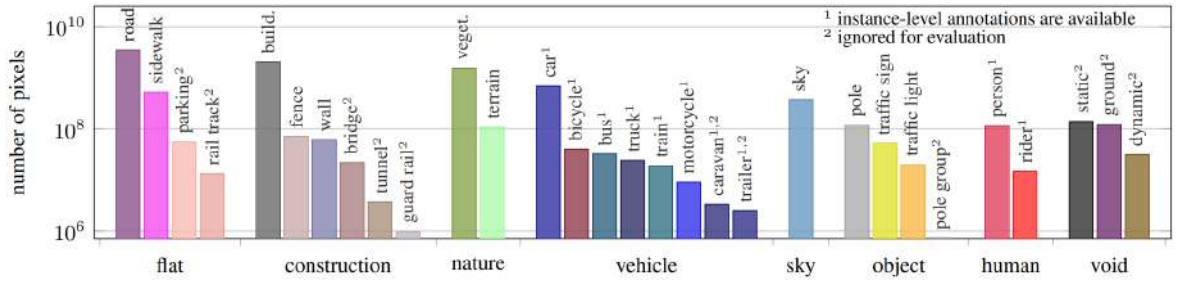

Figure 2.4: **Class taxonomy and distribution of Cityscapes.** Figure is taken from the original source [14].

OODness of an object is defined by the distribution of attributes of the classes the model is trained with. In that sense, the ID dataset and the class taxonomy the model is trained with are of significant importance as they affect the very definition of the problem. We use Cityscapes [14] as the ID dataset and use models that are trained with the Cityscapes training set. Cityscapes contains street-view images from different cities in Germany. Naturally, our ID set consists of 19 classes in the Cityscapes class taxonomy. Figure 2.4 shows the classes and pixel distributions of the Cityscapes dataset, and Figure 2.5 shows example labeled images with the same taxonomy.

Figure 2.5: **Example labeled images from Cityscapes [14] dataset.** Each pixel is color coded by their semantics.

3 Post-hoc Non-invasive Generalizable OOD Detection

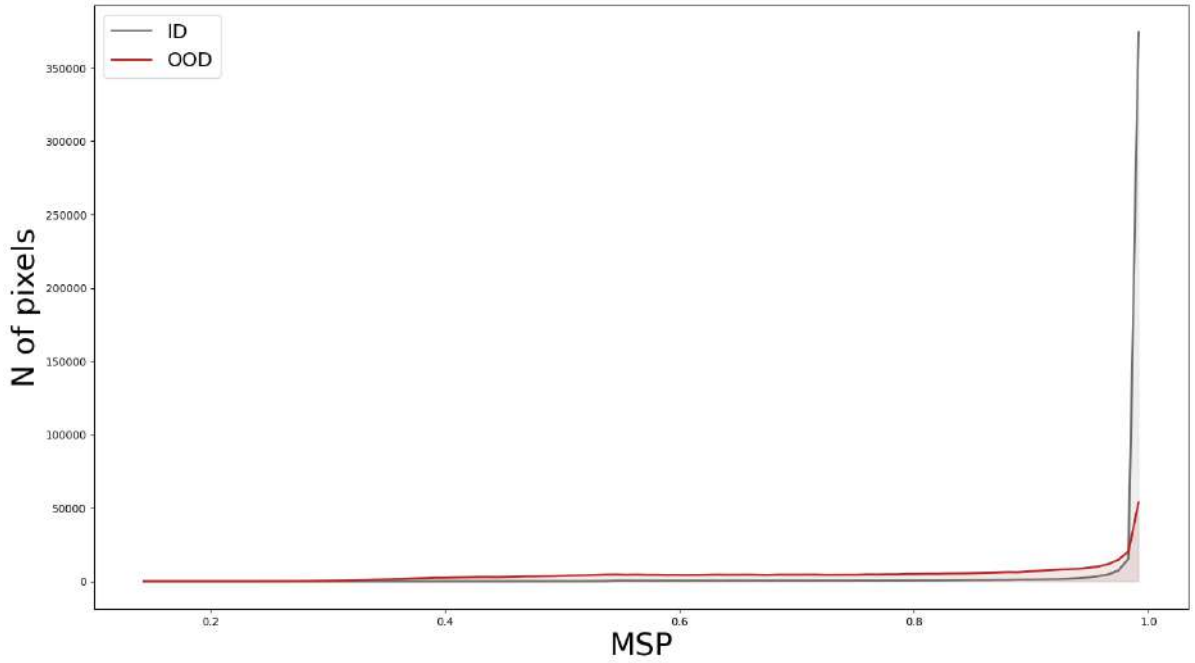

Figure 3.1: **Distribution of the number of ID and OOD pixels under MSP [24].** ID and OOD distributions are not discriminable as an intersection is expected with high-confidence OOD and low-confidence ID pixels.

In this chapter, we review noninvasive, generalizable, post-hoc OOD detection methods to provide an overview of the previous works on the OOD detection domain.

3.1 Max Softmax Probability

OOD detection problems would have a direct solution if the predictive model were ideal. An ideal model would have perfect prediction accuracy and calibration. In other words, it would have a discrete binary prediction that is always correct for any ID object. Since such a model would have no confusion between the samples drawn from ID distributions, any confusion on the model would directly implicate the presence of an OOD. However, as we move further from an ideal model, the expected probability of an ID confusion increases. Without considering the possible ID confusion, one of the baselines for the OOD detection is set as the confidence of the predictions [24]. Even though confidence measures how well an input fits in one of the expected distributions, it fails to discriminate between an ID and an OOD confusion. For a given input x, the confidence of the prediction is calculated as follows:

$$c(x) = max_c p(c|x) \tag{3.1}$$

The expectancy of input as an OOD is negatively correlated with confidence. Thus, a measure for detecting the OOD is inverse of the confidence and calculated as $-max_c p(c|x)$. In the context of OOD detection, using confidences as an OOD measure is called **Max Softmax Probability** in literature [24]. In Figure 3.1, we plot the number of pixels for each confidence value of ID and OOD pixels. We can show that ID and OOD distributions are not discriminable under MSP. Even though most ID pixels have high confidence, the number of pixels with low confidence is not negligible. Interestingly, OOD pixels also tend to have high confidence. Considering the expected number of ID pixels is higher than the OOD pixels, a significant overlap between the low-confidence ID and high-confidence OOD inputs is expected. In segmentation, distant objects and object intersections generally have low confidence. The shortcoming of MSP is mainly falsely recognizing such low-confidence areas as OOD.

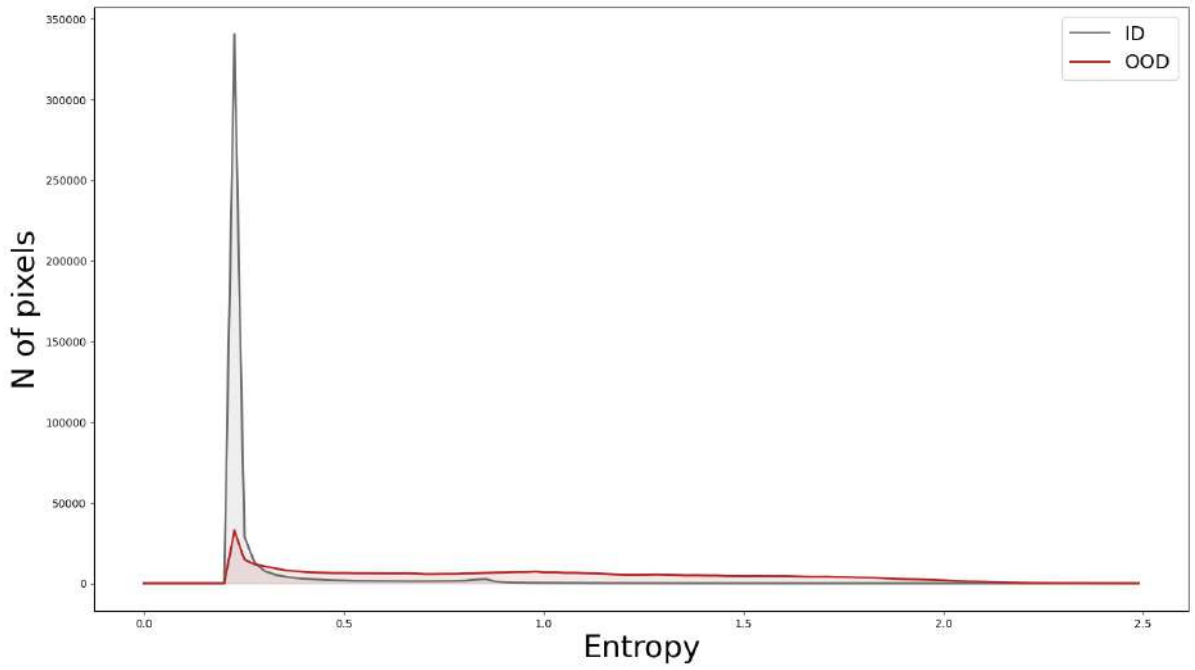

Figure 3.2: **Distribution of ID and OOD pixels under entropy [24].** ID pixels are clustered in the low entropy region. OOD pixels are distributed across high entropy regions.

3.2 Entropy

While having similar confidence values, two predictions can differ in terms of the informative value contained in the confidences of the remaining classes. For instance, confusion between only two classes highly indicates an ID confusion on object intersection, whereas if the confidence is distributed more equally among all classes, prediction offers less information about the input. Considering the distribution of confidences over all classes, **entropy** offers a solution to the shortcomings of MSP [24, 7]. In the posterior domain, entropy for an input x over N classes is defined as:

$$H(c|x) = -\sum_{i}^{N} p(c_i|x) log p(c_i|x) \tag{3.2}$$

In such a context, entropy measures total free information about the input. A discriminative prediction about the input would have low entropy, even if confusion between a few classes is present. In Figure 3.2, we empirically show the distribution of ID and OOD pixels under entropy. Unlike MSP, we now see a lower overlap

between the two distributions. ID pixels are densely clustered in low entropy values, containing higher predictive information. OOD pixels are uniformly distributed across different entropy regions. We still observe an overlap over low entropy regions. However, entropy yields a more discriminable distribution over ID and OOD pixels than MSP. Even though being more discriminative than MSP, entropy still suffers from the closed-world assumption. Entropy is defined on the $p(c|x)$ domain, where for a discriminatively trained network, the closed-world assumption follows that the global set only consists of ID. Overcoming the implications of the closed-world assumption, later methods work on the $p(x|c)$ domain, where no such underlying assumption about the global set is made.

3.3 Max Logits

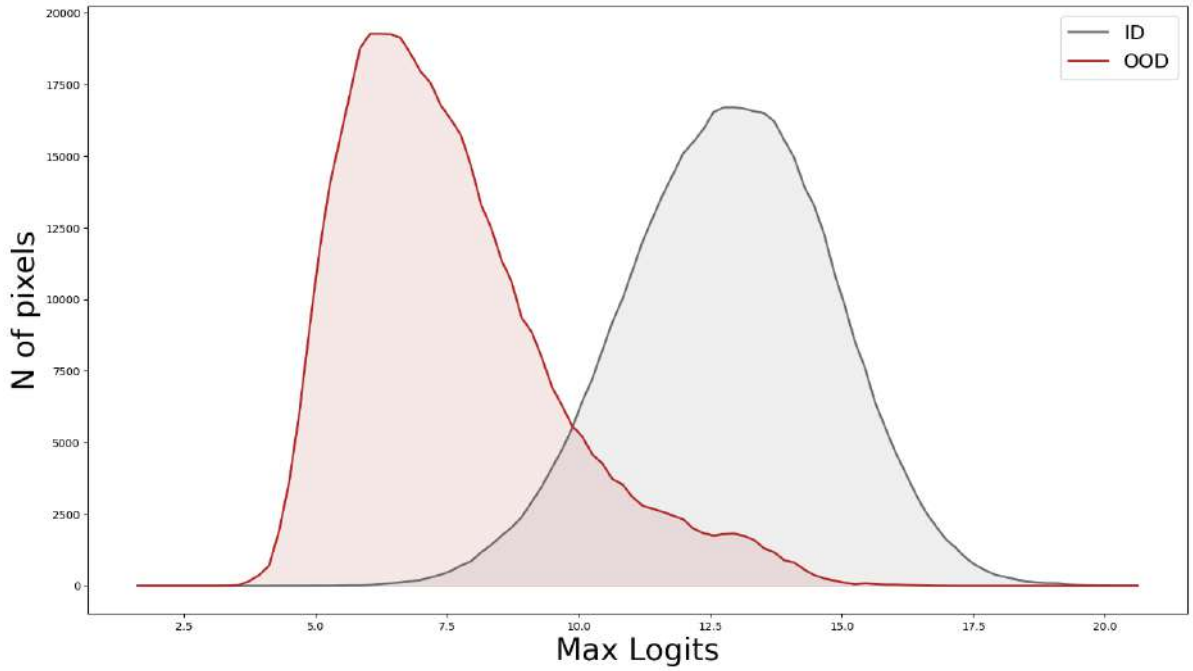

Figure 3.3: **Distribution of ID and OOD pixels under max logits [23].** While having a similar variance, the mean of the max logits of ID and OOD inputs differ.

Logit or class score serves as a density of a class c_i on input x and is independent of the densities of other classes, rendering the logit domain free from the assumption

that the global set only consists of ID set. The closed-world assumption normalizes the total density on x in the confidence domain. As a result, an input with low density can have high confidence if there is an asymmetry in the densities of different classes. On the other hand, one would expect low-class scores regardless of their ranking in an OOD input. Based on these, the maximum of the class scores, or max logits, is often utilized as a measure for OOD detection [23]. The high value of the maximum class score indicates an ID, and the low value of the class score indicates an OOD. Max logits is formalized as the inverse of the maximum class score and formulated as:

$$S_n(x) = -max_c f(w_c, x) \tag{3.3}$$

In equation 3.3, $f(w_c, x)$ is the class score for class c. Note that, measure is also negated on the maximum of class score. Similar to MSP, value of the max class score is also negatively correlated with the expectancy of an input being OOD. In Figure 3.3, we empirically show the distribution of ID and OOD pixels under max logits. Note that the figure shows noninverted values to demonstrate the divergence of the two distributions in the original form. It is apparent that, compared to confidence domain models, max logits, as an ood detection measure, yields more divergent ID and OOD distributions. By leveraging the mentioned properties, recent OOD detection methods commonly use approaches defined in the logit domain. For instance, Standardized Max Logits(SML) [25] normalize the class scores by means and standard deviations calculated from the ID inputs. For class score s_c, they calculate the standardized class scores s'_c as follows:

$$s'_c = \frac{s_c - \mu}{\sigma} \tag{3.4}$$

where μ and σ are the mean and standard deviation of class scores, calculated from the sampled scores from the training set.

3.4 Free Energy

Instead of only focusing on the maximum class scores, "Energy-based Out-of-distribution Detection" [32] claims the Helmholtz free energy on class scores as a measure for OOD detection by explaining the relation between the discriminative objective function and energy-based methods(EBM) [26, 17]. They show the

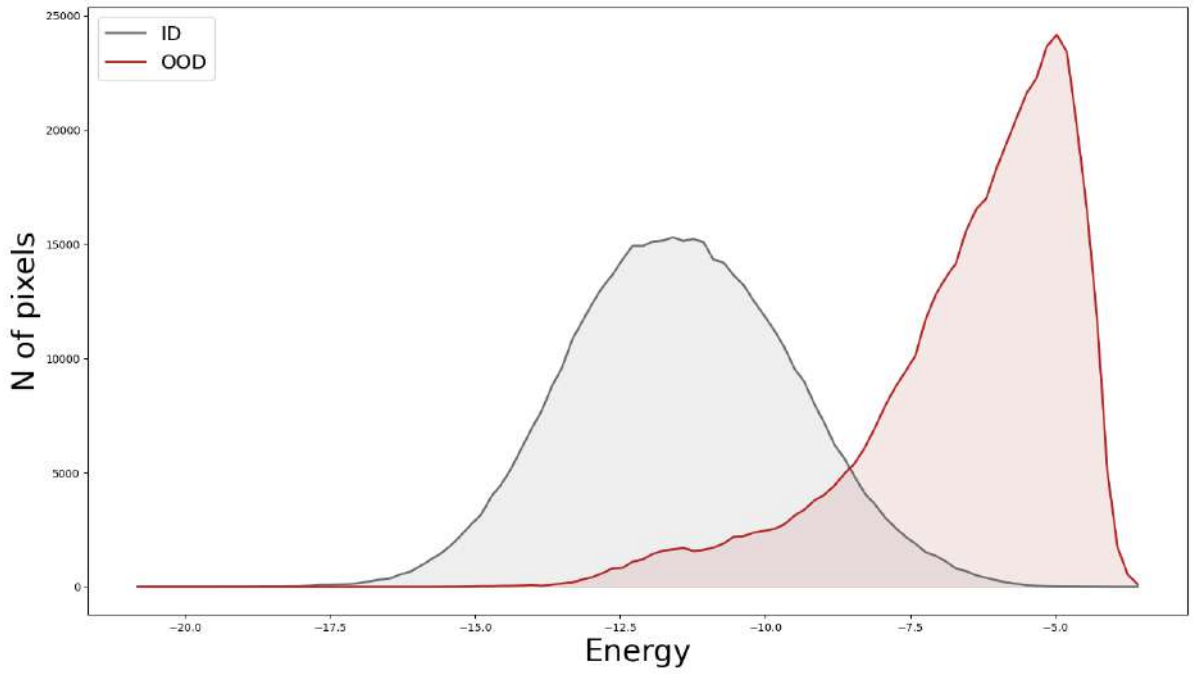

Figure 3.4: **Distribution of ID and OOD pixels under energy function [32].** OOD pixels are heavily clustered in high free energy regions.

interpretation of logits as negative energy function as follows:

$$-E(x,c) = f(w_c, x) \tag{3.5}$$

They show the equivalence of Gibbs distribution of energy values and the softmax function of the discriminative training:

$$p(c|x) = \frac{e^{-E(x,c)/T}}{\int_i e^{-E(x,i)/T}} = \frac{e^{f(w_c,x)/T}}{\sum_i e^{f(w_i,x)/T}} \tag{3.6}$$

Using the interpretation, they define the total free Energy as:

$$S_e(x) = -T \cdot log \int_i e^{-E(x,i)/T} = -T \cdot log \sum_i e^{f(w_i,x)/T} \tag{3.7}$$

Similar to the relationship of MSP and entropy, energy is a measure of class scores over all classes. Formulation estimates the overall density of all classes over an input x, as it can be interpreted as a negative log sum of $p(x|c)$ for all classes. Instead of estimating the probability of input x being a member of a class c, it measures an

estimate parallel to the probability of input x being generated by one of the assumed ID distributions. In this case, ID distributions are classes in the known taxonomy. As estimating if an input x is drawn or generated by one of the ID distributions parallel to the task of OOD detection, the energy function shows competitive results in the domain. As in the max logits, the energy function is also free from the closed-world assumption, as defined on the pre-softmax scores.

Figure 3.4 shows the distribution of ID and OOD inputs under the free energy function. We observe that OOD inputs are densely clustered in high-energy regions. OOD pixels are more densely clustered, thus having a lower standard deviation than the max logits case. We see a similar distribution of ID pixels on max logits and free energy; the main difference comes from the distribution of OOD pixels. As OOD and ID pixels have different means, more densely clustered OOD pixels, with lower standard deviation, yield a more divergent space for the inputs.

3.5 Decoupling Logits

Decoupling methods investigate the effects of different components of the class scores on OOD detection separately. The dynamic components of the class score are the norm and the position or direction of the embedding. We refer to them as dynamic as they are not constant for different input points. Both have a distinctive effect on class scores. Norms of the embeddings serve as a scalar that all class scores are multiplied. While not having a direct impact on the class prediction, it affects the confidence of the prediction. This interpretation of the norms follows the reasoning for temperature scaling for calibration [20], which is also a scalar used to calibrate the confidence and the accuracy. Following the logic, a line of work in the literature claims the norm of the embedding as an OOD detection measure [42, 53]. As norms relate to confidence, they claim that one would expect smaller norms in OOD.

Another line of work emphasizes the significance of the direction of the embedding, as it is directly related to the model's decision [38, 51, 43]. Class scores are defined by the projection or cosine similarity of the embedding to different class vectors on the linear head, and the direction of the embedding is the primary factor in the ranking of class scores of the same input. They claim this cosine similarity as the OOD detection measure, where one would expect a lower cosine similarity to any of the ID classes

from an OOD input. The weighted sum of both similarities is also proposed, claiming both affect OOD detection with different magnitudes [56]. In the next chapter, we investigate the effects of decoupling on the segmentation domain and introduce a detailed formulation.

4 Post-hoc Norm Decoupling in Pixel-wise Segmentation

In this chapter, we discuss the effects of embedding norms on the segmentation model's decision patterns by examining the entropy and confidence distribution in the embedding space. Following our reasoning, we study the effects of post-hoc norm decoupling on existing OOD measures in pixel-wise segmentation. By doing so, we provide insights into the adaptability and extent of the decoupling methods in the segmentation domain.

4.1 Norms as a Degree of Freedom

In the linear classification head score (logit) vector s, where $s = [s_1, s_2, s_3, ..., s_N]_T$, s_i is score for class i and N is the number of classes, is calculated as:

$$Mz + b = s \tag{4.1}$$

M is the matrix that corresponds to linear head in shape $[CxN]$, C is the number of classes, N is the dimension of the embedding space. z is the embedding vector of the pixel in shape $[Nx1]$, and b is bias. For the rest of the calculations, we will omit bias for simplicity, as it is elementary to also consider bias in the following calculations.

A class score vector of a maximum entropy in the embedding space can be defined as $v = [v_1, v_2, v_3, ..., v_N]_T$ where $v_i = v_j$ for any i, j pair. The confidence vector is uniform where all class scores are equal, and entropy has an upper bound where the confidence vector is uniform [46]. Substituting to 4.1, we can define an equation for an embedding on a maximum entropy point as:

$$Mh = \alpha v_I \tag{4.2}$$

where h is a maximum entropy point and v_I is a vector with all elements equal to 1 and α is a scalar in $\mathbb{R}$. $\alpha v_I = v$, means that set of αv_I and v are equal. Equation 4.2 has infinitely many solutions for h if $rank(M) = C$. We can assume that M follows the rank constraint since we expect a fully trained network to be able to successfully discriminate between C classes. h having infinitely many solutions shows us that there are infinitely many maximum entropy points in the embedding space.

Proposition 1. *The confidence vector of an embedding is dependent on the relative position of the embedding to maximum entropy points, regardless of the magnitudes of the class scores.*

Proof. Relative position of an embedding to a maximum entropy point can be represented by a vector from the maximum entropy point to the embedding. We can rewrite any embedding as a sum of a maximum entropy point and its relative distance to the maximum entropy point:

$$z = h + k \tag{4.3}$$

For the relative distance k and maximum entropy point k:

$$\begin{aligned} s &= Mz \\ s &= M(h+k) \\ s &= Mh + Mk \\ k' &= Mk \\ s &= \alpha v_I + k' \end{aligned} \tag{4.4}$$

Note that M is constant, which makes k' identical for any h with the same k. For our current score vector s', confidence for any class prediction i is defined as:

$$\begin{aligned} p_i &= \frac{\exp(s_i)}{\sum_j \exp(s_j)} \\ p_i &= \frac{\exp(\alpha + k'_i)}{\sum_j^N \exp(\alpha + k'_j)} \end{aligned} \tag{4.5}$$

where k'_i and s_i are the i^{th} component of k' and s respectively. To show that the class prediction p_i is independent of the magnifying factor α, we differentiate p_i with respect to α using the quotient rule. We rewrite the Equation 4.5 and differentiate as following:

$$\frac{d}{d\alpha}p_i = \frac{d}{d\alpha}\frac{u}{v} \tag{4.6}$$

where u is equal to:

$$u = \exp(\alpha + k'_i) \tag{4.7}$$

and v is equal to:

$$v = \sum_{j}^{N} \exp(\alpha + k'_j) \tag{4.8}$$

Following the quotient rule, we rewrite Equation 4.6 as:

$$\frac{d}{d\alpha}p_i = \frac{(\frac{d}{d\alpha}u)v - (\frac{d}{d\alpha}v)u}{v^2} \tag{4.9}$$

Since both u and v are exponents, $\frac{d}{d\alpha}u = u$ and $\frac{d}{d\alpha}v = v$. Substituting into Equation 4.9:

$$\begin{aligned} \frac{d}{d\alpha}p_i &= \frac{uv - uv}{v^2} \\ \frac{d}{d\alpha}p_i &= 0 \end{aligned} \tag{4.10}$$

□

We can prove that infinitely many points with different class scores in the embedding space have the same resulting confidence vector p. Indeed, we can rewrite

any point in the embedding space as a sum of a maximum entropy point h and a vector k as $z = h + k$. If we add vector k to any other maximum entropy point h', we get another point $z' = h' + k$. By equation 4.10, since two vectors h and h' are both maximum entropy points, resulting confidence vectors of both points z and z' would be equal.

Let us define an example embedding $z_e = h_e + k_e$ with the assumption $\|z_e\| > \|k_e\|$. Origin is a maximum entropy point with the assumption $b = 0$. As a side note, if we were to include the origin, maximum entropy points would simply shift by $-b$. Since the origin is also a maximum entropy point, $z'_e = 0 + k_e$ has the same confidence vector as z_e. We know that $\|z'_e\| = \|k_e\|$ thus $\|z'_e\| < \|z_e\|$. Note that we can repeat selecting different maximum entropy points infinitely, getting an infinite set of embeddings with the same confidence vector and differing norms. Further, if we define an additional move from any of the embeddings, the change in the confidence vector is identical for any vector in the said infinite set. We can write the additional move as $x_e = h_e + k_e + m_e$. And again selecting origin as the reference basis, $x'_e = 0 + k_e + m_e$ with x'_e and x_e having the same confidence vector. One can aggregate the move to any point with the same confidence vector, and the resulting confidence vector would be the same, as the sum of two move vectors can be represented as a single move. As a result, the gradient of the objective function with respect to different embeddings with the same resulting confidence vector would be the same, as the same move on the embeddings would result in an identical change in the confidence vector, resulting in an identical gradient field around all points. Note that the same gradient field repeats itself infinitely with differing norms.

By recalling the softmax function, our findings parallel the observation that estimating class scores for the model is an underconstrained problem. For the same $p(c|x)$, there are infinitely many solutions for $f(w_c, x)$. There is no incentive for the model to converge to any magnitude of the class scores, given that no additional regularization is used. Decision-wise, selecting any arbitrary maximum entropy point does not impact the objective function as α in Equation 4.5 is a degree of freedom decoupled from the resulting confidences. The ambiguity of α aggregates itself to the class scores primarily via the norm of the embedding. Being primarily related to a degree of freedom decoupled from the model's decision mechanism, one can not assume a different distribution of norms between ID and OOD inputs.

We discuss the overall magnitude of the class scores, and as an extension, the norms of the embeddings are decoupled from the model's decision mechanism. The softmax score normalizes class scores for the same input, and the decisive factor is their ranking and relative magnitudes. This discussion calls for the question: *What information do the norms of the embeddings contain?*

4.2 Logits Under External Shift

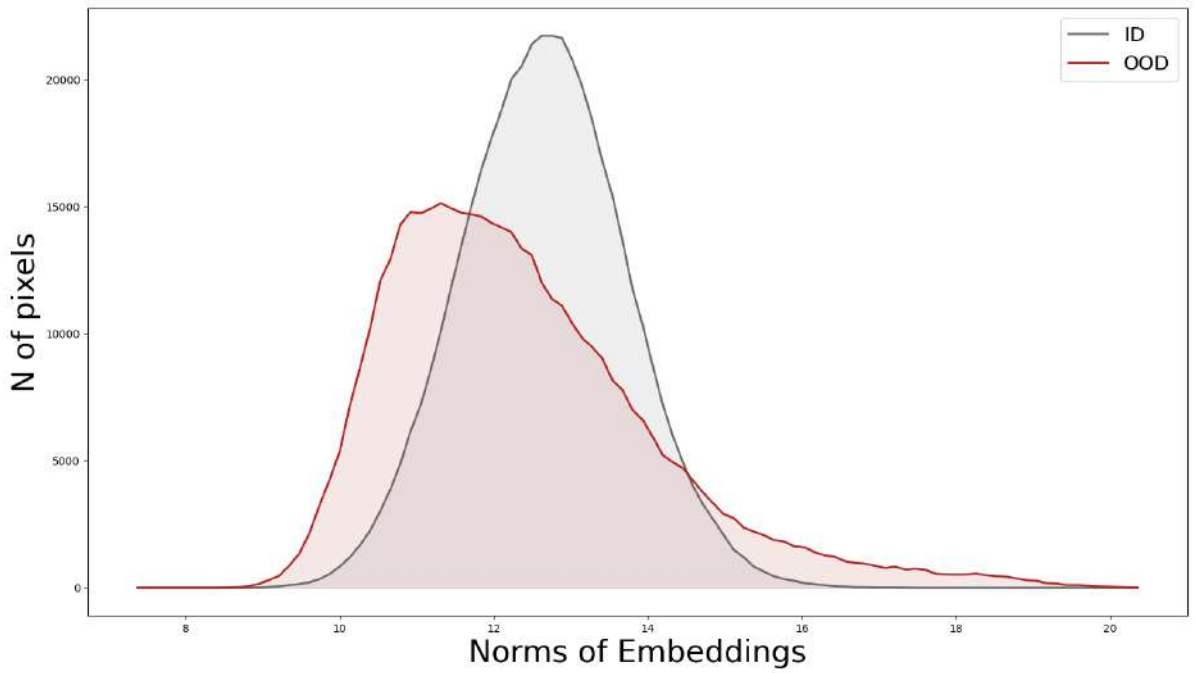

Figure 4.1: **Distribution of the norms of ID and OOD embeddings.** A high intersection between ID and OOD supports our claim for ambiguity.

The external shift is the shift in the distribution of external factors, including color density, distance, brightness, and noise. It can also be phrased as a shift in the input distribution. A model trained with objects varying in external factors must stay robust under different external shifts. This need calls for a degree of freedom in the class scores and embeddings decoupled from the model's decision mechanism. We claim that the norm of embeddings primarily contains information on the distribution of external factors. In other words, the model uses norms as a balancer to stay robust under external shifts. Said external factors are decoupled from the discriminative

decision for the model yet still contain information about the input pixel. The model can use decoupled degrees of freedom to adjust and normalize the class scores under different external distributions during the training. However, on OOD detection, one expects a similar distribution of external factors over ID and OOD objects in the open-world segmentation domain. Such external factors are not a discriminative measure over the object's OODness, as the main discriminative factor is the whatness.

OOD detection on class scores requires a reliable comparison or ranking between input pixels. Pixels under different external distributions can have varying class scores regardless of the prediction's confidence. This arbitrariness also holds for ID and OOD pixels, as no assumption on external distribution that would directly discriminate the OOD pixels can be made in the open-world segmentation domain. For instance, an OOD with low noise and under dense illumination can have higher class scores than an ID object under shadow or in the distance. Figure 4.1 shows the distribution of norms of ID and OOD pixels. A high intersection between two distributions supports our claims. Also, it is essential to note that the density of the larger values of the norm is higher on OOD pixels.

The discussed effects of the norms on the class scores can be mitigated by projecting embeddings into a hypersphere where the norms are constant. In the resulting embedding space, class scores are dependent on the cosine similarity to class vectors on the linear head, as well as the class weights. This way, the unpredictability of class scores that is imposed by external factors is eliminated. Note that this can be performed post-hoc as following:

Returning to the equation 4.1, let us define the logit for a single class as:

$$s_i = M_i z + b_i \tag{4.11}$$

where s_i and b_i are score and for the i^{th} class respectively. Similarly, M_i is the i^{th} row of the linear head corresponding to the respective class. We will omit the bias, as our empirical findings suggest that bias is negligible. Projecting the embedding into a hyperspace is equivalent to dividing it by its norm. The resulting normalized embedding is calculated by:

$$z' = \frac{z}{\|z\|} \tag{4.12}$$

Substituting z with z', the class score on the hypersphere becomes:

$$s_i' = M_i \frac{z}{\|z\|} \tag{4.13}$$

We can rewrite the M_i as:

$$M_i = \omega_i m_i \tag{4.14}$$

where m_i is the normalized linear head vector with norm ω. Substituting the equations 4.14 and 4.12 into equation 4.13, we formulate the normalized class score as:

$$\begin{aligned} s_i' &= \omega_i m_i \frac{z}{\|z\|} \\ s_i' &= \omega_i m_i z' \\ s_i' &= \omega_i cos(\alpha_i) \end{aligned} \tag{4.15}$$

In equation 4.15, $cos(\alpha_i)$ is the dot product between two normalized vectors m_i and z', which is equivalent to the cosine similarity between the embedding and the class vector.

We perform an experiment on ID images to illustrate the relation between norms and external factors and the effect of normalized class scores. In this experiment, we sample a subset of ID images from cityscapes and impose an external shift. We reduce the color intensity of the images and apply a Gaussian blur. In the top graph in Figure 4.5, we compare the distribution of norms of the same images with and w/o external shift. There is a distinctive shift in the distribution of norms due to the added external shift. We plot the class scores on the middle graph with a highly parallel shift to the norms. To support our claim that the information about the external distribution is primarily maintained in the norm, we decouple the norm from the equation and show the distribution of normalized maximum class scores on the bottom graph. Two distributions being nearly indistinguishable shows that one can remove the external shift in the class scores by normalization. In other words, normalized class scores are robust to external factors other than objects' whatness, which is a highly sought-after property in OOD detection.

We emphasize that our insights about the norms of the embeddings only hold for

domains with the assumption of a similar external distribution between ID and OOD inputs. Thus, norms can remain a discriminative measure in a domain with a direct difference of external distributions between ID and OOD is expected.

As we establish the motivation on why norms of the embeddings are not a discriminative factor and cause ambiguity on class scores in Section 4.1, and how such ambiguity can manifest itself in a practical sense in this section, we move on to using attained robust representation, normalized class scores, in OOD detection.

4.3 Decoupled Post-hoc Energy

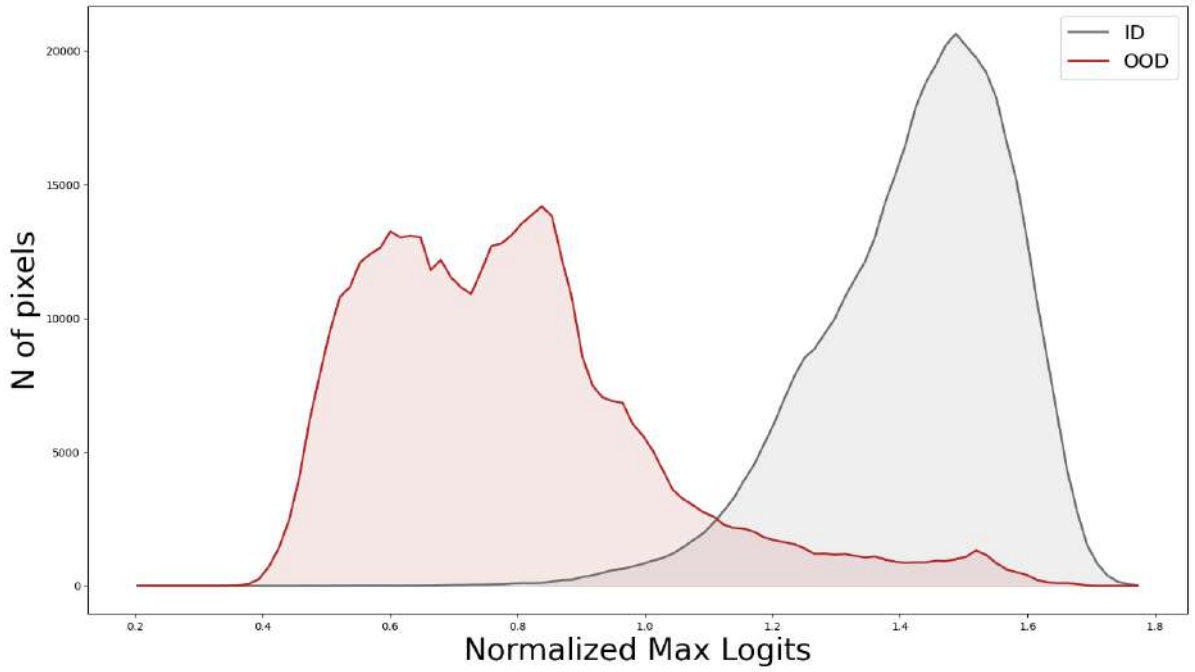

Figure 4.2: **The distribution of ID and OOD pixels under normalized max logits.** We observe an increase in divergence in ID-OOD distribution when normalization is applied to class scores.

Following our motivation, we extend normalized class scores to score-based post-hoc OOD detection methods. Notice that normalizing class scores does not hinder the properties of non-invasiveness and generalizability. Normalizing does not require any change in the model structure or retraining and also does not change the model's prediction accuracy. The reason for unchanging accuracy is that each class score

vector is divided by the same norm value, which does not change the ranking between class scores. Also, norm normalization can be applied in any discriminatively trained segmentation model, thus retaining the generalizability of the class score-based post-hoc methods.

Returning to Equation 4.15, which shows the normalized class score for a single class, we can define the vector of normalized class scores for input x as:

$$s' = \omega cos(\alpha) \tag{4.16}$$

where s' is the vector of the class scores, ω is the vector of the class weights, and α is the vector of the cosine similarity. Notice that the only difference between initial and normalized class scores is the division by the norm. Normalized class scores can also be represented by class scores divided by the norm:

$$s' = \frac{f(w, x)}{\|z\|} \tag{4.17}$$

Following this, we can define an OOD score Normalized Max Logits:

$$S'_n(x) = -max_c \frac{f(w_c, x)}{\|z\|} \tag{4.18}$$

Figure 4.2 shows the ID and OOD input distributions under Normalized Max Logits. Recalling the distribution in Figure 3.3, we observe less intersection on the Normalized Max Logits than on the Max Logits. Increased divergence yields better OOD detection. The change in the distribution by normalization is parallel to our claim of arbitrariness aggregated to class scores over norms.

The Energy function is an OOD score measure that is defined by the class scores. Empirical findings of the literature suggest that the Energy function is a better estimate for OOD compared to Max Logits. We employ normalized class scores with

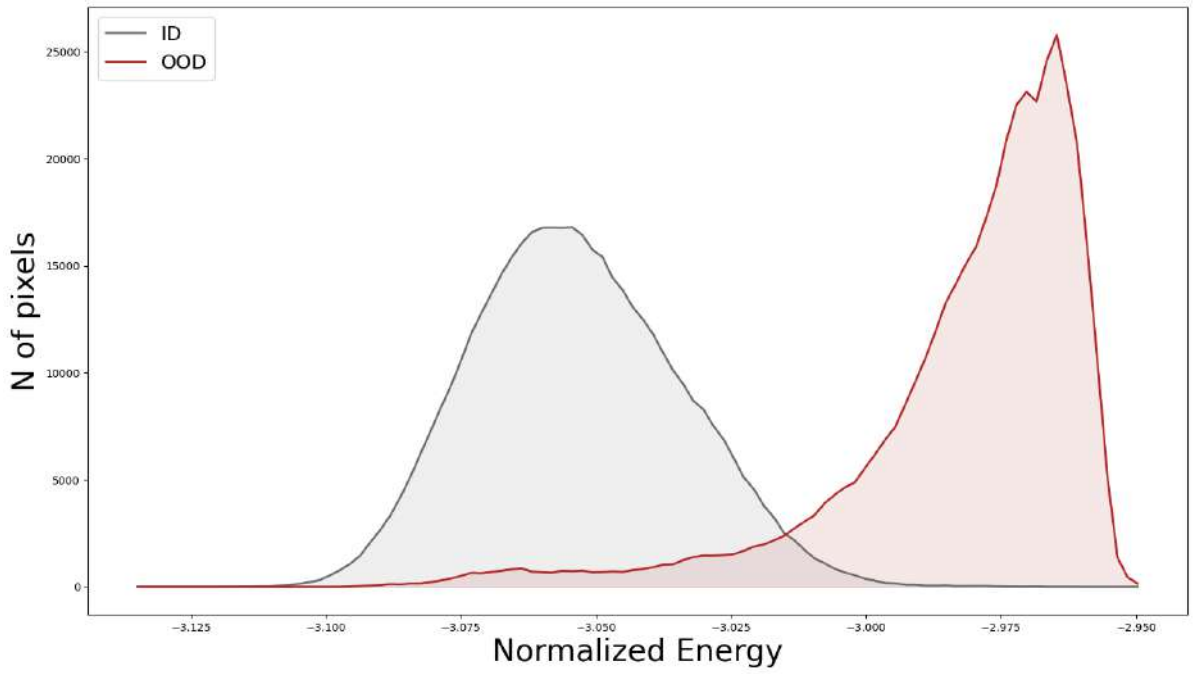

Figure 4.3: **The distribution of ID and OOD pixels under normalized energy function.** Increased divergence and decreased intersection between ID and OOD pixels is empirical evidence for the effectiveness of normalized class scores with energy function [32] in OOD detection.

the energy method by redefining the energy function as:

$$-E'(x) = \frac{f(w_c, x)}{\|z\|} \tag{4.19}$$

By substituting to the total Helmholtz free energy, we define the normalized energy function as:

$$S'_e(x) = -T \cdot log \sum_i e^{\frac{f(w_i,x)}{\|z\|T}} \tag{4.20}$$

In practice, we use a temperature of 1. By substituting T as 1, in a simplified form, the normalized energy function is:

$$S'_e(x) = -log \sum_i e^{\frac{f(w_i,x)}{\|z\|}} \tag{4.21}$$

Figure 4.3 shows the ID-OOD distribution under normalized energy. Compared

to energy without normalization, ID pixels retain a similar distribution. However, we observe a higher density of OOD pixels in high-energy regions. Consequently, the total ID-OOD intersection is reduced. As a result, one expects a higher OOD detection accuracy on normalized max logits. Our empirical findings in Section 5 is parallel to our observation.

4.4 Class Weights in Segmentation

Learned class weights in the linear head are a learned weighting of ID class distributions. In optimization, the model is given the freedom to adjust cosine similarities between embeddings and the linear head vectors, also considering the class weights.

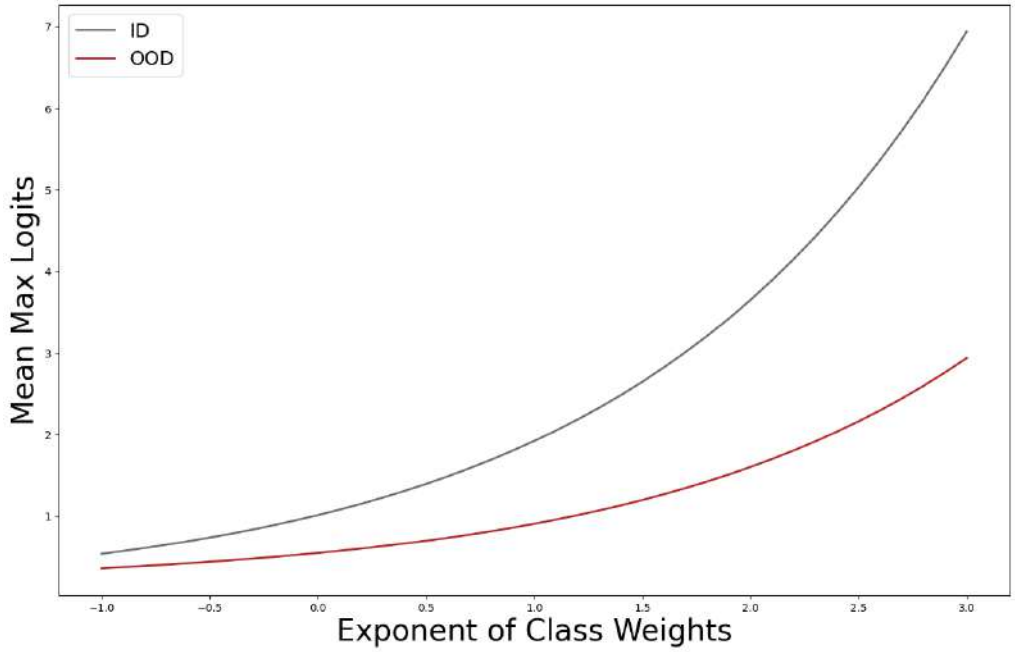

Figure 4.4: **Mean value of max logits [23] with different η values.** ID and OOD distributions further diverge as we increase η.

Classification methods that use decoupling suggest normalizing by class weights [38, 56]. In the segmentation domain, our empirical findings suggest otherwise(Chapter 5 Figure 5.7). Normalizing by class weights, in our case, reduces the divergence between ID and OOD distributions and, as an extension, OOD detection accuracy. On the contrary, increasing the class weights further penalizes the embeddings with

low cosine similarity to any class vector. We propose introducing exponent η on class scores as a tunable parameter:

$$s'_w = \omega^{\eta} cos(\alpha) \tag{4.22}$$

With the introduction of updated class scores, the energy function becomes:

$$S'_w(x) = -log \sum_i e^{\frac{\omega^{\eta-1} f(w_c, x)}{\|z\|}} \tag{4.23}$$

Figure 4.4 plots the mean class scores of ID and OOD inputs under different η values. Notice that $\eta = 0$ is equivalent to normalizing the class weights. We show a correlation with η and divergence of ID and OOD distributions. While normalizing by class weights brings the two distributions closer, increasing it further diverges.

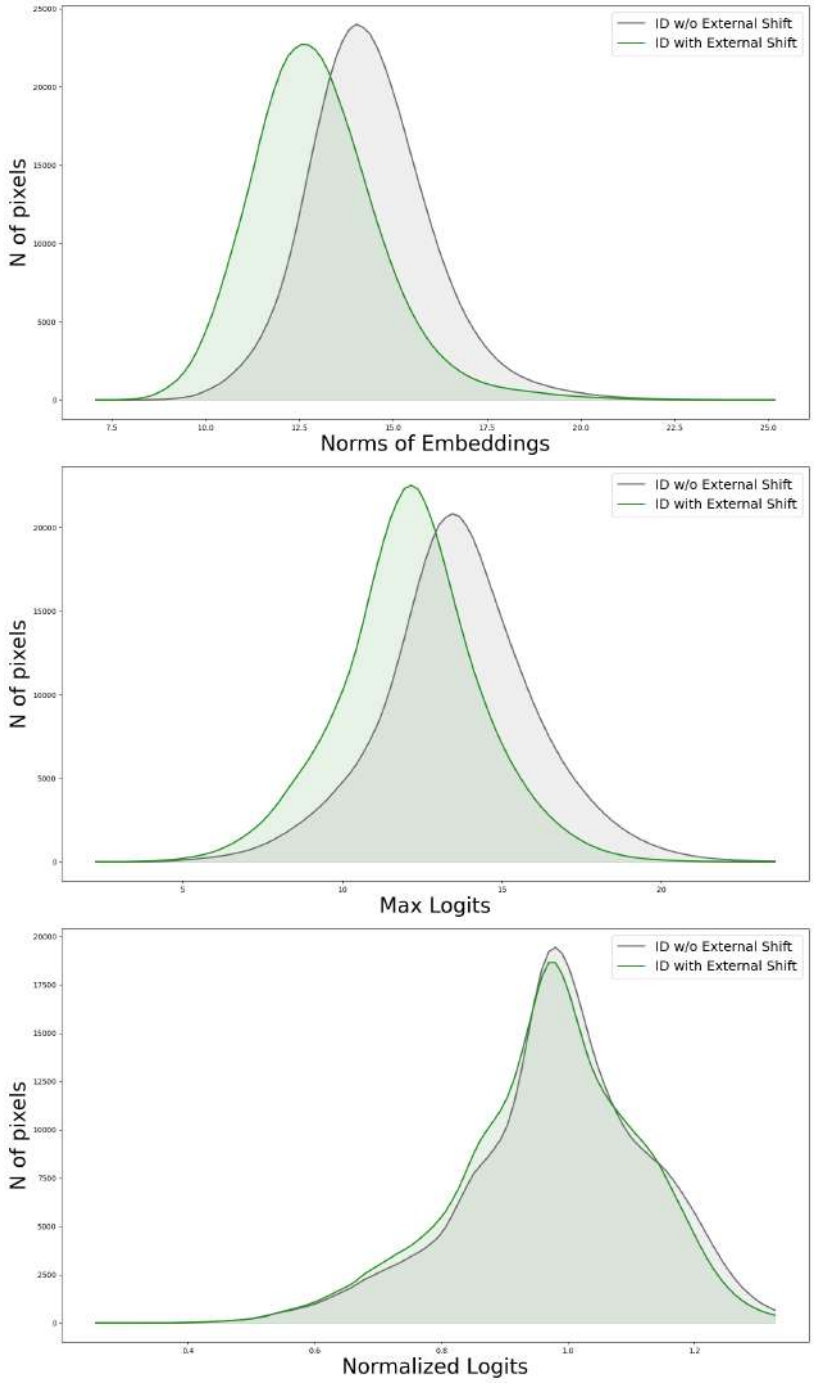

Figure 4.5: **Experiments of maximum class scores under domain shift.** The top graph shows the distribution of norms of ID inputs with and w/o an external shift. The middle graph shows the distribution of maximum class scores under the same constraints. The bottom graph shows recovered logit distributions with normalization.

5 Evaluation

In this section, we present a thorough evaluation of the performance of OOD detection methods employed with norm normalization. We compare our methods against the previous baselines and show the effect of the normalization in different OOD detection methods. To demonstrate the robustness and consistency of our methods, we test multiple models on multiple datasets. We quantitatively evaluate our methods with three different OOD accuracy metrics. We include a discussion of underlying reasons for the performance of visual transformer methods on OOD detection. Later, we present a qualitative analysis by showing heat maps of the OOD scores, demonstrating the direct effect of score normalization. We follow up with additional experiments where we further investigate the effect of normalization and class weights. Lastly, we discuss the limitations of our method.

5.1 Metrics

We follow the literature to examine the accuracy of our methods for OOD detection [24]. In doing so, we use commonly used metrics in pixel-wise and instance-wise OOD detection, such as AP, AUROC, and FPR95. Said metrics measure the divergence or discriminability of the scores of OOD and ID inputs.

In pixel-wise detection tasks, **True Positives(TP)** are the number of pixels in the target that are successfully detected. **False Positives(FP)** are the number of pixels outside the target that are wrongfully detected as a target, and **False Negatives(FN)** are the number of pixels on the target that are not detected.

Following the definitions, **Precision** is the ratio of correctly predicted targets to the total number of predictions. **Recall** is the ratio of correctly predicted pixels to the

total number of targets. Both calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Average Precision(AP) is generally defined as the area under the Precision-Recall curve. This metric is standard in evaluating detection-based tasks. In OOD detection, AP is calculated for different confidence thresholds of OOD scores. AP is averaged across different thresholds based on the TPR(True Positive Rate) scores. Averaged AP score across different thresholds can be formalized as:

$$AP = \frac{1}{N} \sum_{t>\gamma} AP_t \tag{5.3}$$

where t is the confidence threshold, AP_t is the Average Precision at threshold t and γ is a lower bound for thresholds set based on TPR values.

AUROC is the area under the TP-FP curve. TP-FP curve embeds the TP rates under different FP rates. AUROC measures the accuracy of the model under different thresholds of FP rates. By doing so, it plots how TP rates change under increasing FP rates. The model would have a TP of 1.0 regardless of the FP rate in an ideal scenario. Such a scenario results in a value of 1, which is the maximal for AUROC. The worst-case scenario is a TP of 0 under different FP rates, which results in an area of 0 under the TP-FP curve. AUROC is scaled between 0-1, and a higher value indicates a better OOD detection accuracy.

False positives cause critical disturbances while employing OOD detection methods in real-world settings. To measure the safety of the method, FP rates at high TP rates are of significant importance. **FPR95** is the FP rate for thresholds in which $TPR > 0.95$.

5.2 Thresholding

Thresholding on OOD scores is typically done as follows:

$$L(x) = \begin{cases} 1 & \text{if } S(x) \geq \gamma \\ 0 & \text{if } S(x) < \gamma \end{cases} \tag{5.4}$$

where $L(x)$ is a binary mask for OOD objects, $S(x)$ is OOD score and γ is the threshold [16].

In the distribution graphs we show, one can select the threshold in the intersection of two distributions to maximize the metric scores. However, in the practical case, the threshold is determined by the objective of the application. In a safety-critical deployment, one can select a threshold that would set the FP rate to 0. In a domain where the importance of detecting OOD outweighs the false positives, one can choose a threshold that would result in a TP rate of 100. Selecting threshold is a domain-specific problem. It also presents an equivalent challenge to all OOD score methods. For that reason, OOD detection metrics measure the performance under different thresholds determined by TP and FP rates, leaving selecting a specific threshold out of the evaluation.

5.3 Datasets

We use the datasets under the **Segment Me If You Can(SMIYC)** [6] benchmark to evaluate our methods. SMIYC baseline offers different datasets with different OOD scenarios and varying domain shifts for obstacle and anomaly detection in semantic segmentation. These datasets contain images with OOD objects from varying sources, such as street-view images and online sources. SMIYC provides pixel-wise labeling for the validation images with a taxonomy of ID-OOD. Images are binary-labeled following the taxonomy. The classes of the cityscapes taxonomy define the ID for SMIYC. Any object outside the cityscapes taxonomy is defined as an OOD. While testing with the SMIYC datasets, the objective is to generate a class score for each pixel, a higher score indicating a higher probability of being an OOD.

The first dataset we use for the evaluation is Lost and Found. It contains different street-view images with OOD objects on the road, selected from different categories to create a diversity of OOD objects. The *Standard Object* category contains boxes

Figure 5.1: **Example Images on Lost and Found [6].** Lost and Found poses a challenge for OOD detection under low domain shift.

and crates of different colors. The *Random Hazards* category contains varying outdoor and indoor objects, including pylons, tires, plastic bags, and bumpers. The *Emotional Hazards* category contains critical objects such as dogs, balls, and bobby cars. Lastly, the *Humans* category contains kids.

The Lost and Found dataset employs a relatively low domain shift, as images are taken from daylight street-view images similar to the Cityscapes ID images. However, objects are placed at differing distances, introducing the challenge of detecting OOD in a small field of view. Figure 5.1 shows example images sampled from the Lost and Found dataset. As seen in the example images, objects are placed on the street-view scenes manually. In the left image, the green crate is OOD, and an OOD detection method should give high scores for only pixels that correspond to the crate. In the right image, multiple OOD objects are placed at the scene at the same time, thus creating a challenge for dense multi-instance OOD detection. We use publicly available test split of Lost and Found to evaluate our methods.

Unless we state otherwise, the ID-OOD distribution graphs we show are sampled from the Lost and Found dataset.

The second dataset we use is Anomaly Track. It contains various images sampled from online sources, so it employs a high domain shift for a model trained with Cityscapes images. Also, it contains a wide variety of OOD classes as. Figure 5.2 shows two example images for the Anomaly Track dataset, which shows an example of a high shift from the external assumptions of the Cityscapes. Both images are focused on a single instance with varying backgrounds. Since images are not street-view images, the Anomaly Track tests the accuracy of OOD detection in arbitrarily composed input images. For our evaluation, we used the validation set of Anomaly

Figure 5.2: **Example Images on Anomaly Track [6].** Anomaly Track contains images from online sources, extending the OOD detection challenge to different image compositions.

Figure 5.3: **Example Images on Obstacle Track [6].** Contains various OOD objects placed on the road in rural areas.

Track.

The third dataset we use is Obstacle Track. Like Lost and Found, it consists of OOD objects on the road. Different from Lost and Found, it does not contain street-view images from populated cities but from rural areas. Although it does not have the same street-view structure as the Cityscapes dataset, it employs a lower shift than the Anomaly Track. In Figure 5.3, we show example images from the Obstacle track. The left image contains multiple of the same object with different colors, creating a medium for testing the OOD detection scores with changing external factors. The right image contains a dog as OOD object. The SMIYC benchmark generally contains many images with animals as OOD, emphasizing the importance of detecting critical OOD objects with an emotional and ethical factor. We use the validation set of Obstacle Track.

The last dataset we use is Road Anomaly, a primitive version of the SMIYC benchmark. It contains a mixture of OOD objects placed on the road and images

Figure 5.4: **Example Images on Road Anomaly [6].** As a primitive version of the SMIYC [6] benchmark, Road Anomaly poses a complex challenge in varying conditions.

Method	Lost and Found			Anomaly Track			Obstacle Track			Road Anomaly		
	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR
Norm - N	35.99	0.36	99.99	41.11	12.72	97.08	27.73	0.43	99.99	51.40	10.82	95.80
Max Softmax [24]	93.35	23.08	31.85	81.74	46.27	57.38	70.21	1.94	66.25	70.75	20.56	68.91
Max Logit [23]	95.60	58.95	29.77	85.32	58.46	55.46	74.35	3.94	70.77	76.43	25.88	65.81
SML [25]	90.01	33.63	55.41	71.71	37.27	81.62	47.71	0.63	92.78	63.03	13.67	80.69
Entropy [24]	94.08	41.99	31.63	83.23	54.29	57.11	71.60	3.95	66.67	72.09	23.08	68.76
Energy [32]	95.63	60.07	29.77	85.48	58.17	54.97	75.60	4.80	70.83	77.19	25.77	65.46
N-Max Logit	97.45	71.1	**13.28**	95.18	77.30	**18.63**	86.15	9.57	40.12	83.61	36.33	62.17
N-Energy	96.44	75.53	23.58	95.00	78.17	22.00	95.87	26.49	14.65	86.63	42.78	**58.43**
N-Energy + CW	**97.54**	**83.19**	13.93	**95.27**	**81.28**	22.78	**97.46**	**72.54**	**8.87**	**86.85**	**48.98**	60.97

Table 5.1: **OOD accuracy of normalization.** We present the quantitative analysis of normalization with DeepLabv3+[10]. The N—prefix adds normalization to the method. CW means optimized class weights. We experiment on four different datasets with three different metrics.

from online sources with OOD objects. Because of its variety, it stays challenging and relevant to this date and is used as a benchmark in recent literature. Figure 5.4 shows two examples from the said mixture. On the left, similar to Lost and Found and Obstacle Track, three OOD objects are placed manually on the road. On the right is an online image with an animal as OOD. We use the Road Anomaly test set.

5.4 Quantitative Experiments

Table 5.1 presents the quantitative evaluation of normalized class scores with DeepLabv3+. We use a flat model trained on the Cityscapes train set with a discriminative softmax

loss. We test each method with the same backbone to compare different methods fairly. We do not use any postprocessing, including SML.

The first baseline Norm-N represents the norms as an OOD score with the negative correlation assumption. The negative correlation assumption here follows that a smaller norm indicates a higher score for OOD. We follow this assumption because the literature on norms that claims norms as a distinctive feature of OOD detection follows the same assumption [42, 53]. The negative correlation assumption indicates that the norm is related to the class scores and confidence and, thus, is higher for ID. We show that, for all four datasets, the norm performs the worst. The average precision being near 0 suggests that there is no divergence between ID and OOD distributions under the norm as an OOD score. For the norms, for Anomaly Track and Road Anomaly, AUC or AUROC is close to 50, supporting our claims for a similar distribution of norms between ID and OOD. $AUROC = 50$ indicates that TP and FP rates are equal for all thresholds.

We show the effect of the score normalization on max logits and energy scores. N-Max Logit indicates max logit on normalized scores. Similarly, N-Energy indicates the energy score on normalized scores. N-Energy + CW is the normalized energy score with an increased class weight with $\eta = 1.5$. Normalized max logit outperforms the max logit by **+1.85/+9.86/+11.8/+7.18** for AUC on Lost and Found, Anomaly Track, Obstacle Track, and Roan Anomaly, respectively. With the same order of the datasets, for Average Precision, gain by normalizing class score over max logit is **+12.15/+18.84/+5.63/+10.45**. For FPR, lower scores indicate a better result. FPR reduces by **-16.49/-39.93/-30.6/-1.3** with norm normalization. For max logits, for all metrics on all datasets, there is a consistent improvement with normalization.

Moving on to the energy function as an OOD score, we observe a significant improvement parallel to the max logits results. For AUC, normalizing scores results in a gain of **+0.81/+9.52/+20.27/+9.44** for four datasets respectively. For Average Precision, gain by normalizing is **+15.46/+20.0/+21.66/+17.01**. We also observe a significant decrease in FPR by **-6.19/-36.97/-56.18/-7.03**. Consistent improvement in the energy function supports our claim that normalized class scores can be employed with different OOD measures.

The best-performing method on all datasets is the class-weighted normalized energy score, which outperforms normalized energy and normalized max logit separately

in 10 out of 12 experiments. The two experiments where increased class weight reduces the performance are FPR experiments. However, a significant improvement over the Average Precision outweighs the possibility of a slight increase in FPR. For all four datasets, gain on Average Precision by normalizing the class weights are **+7,66/+3.11/+46.05/+6.00**. Our empirical findings on class weights support our claim that increased class weight further pushes ID and OOD distributions apart. In some cases, said push can result in two distributions passing a threshold with high intersection, resulting in a significant increase in Average Precision. Obstacle Track is an example of such improvement.

In Table 5.2, we show the evaluation of our methods with OCRNet. We follow the same experimental setup with DeepLabv3+.

We observe that the Norm-N score function results with the worst accuracy, suggesting a similar non-divergent ID-OOD distribution with the norm. Similar to DeepLabv3+ experiments, on OCRNet, score normalization yields a consistent increase in the OOD detection accuracy with all metrics on all datasets. On the AUROC, normalized max logits outperform max logits by **+1.66/+18.37/+3.39/+11.74** respectively for four datasets, with the most significant improvement on Anomaly Track. On Average Precision, the gain by normalizing class scores is **+15.41/+34.7/+27.1/+11.74**, significantly improving over the Anomaly Track and Obstacle track. FP rate reduces by **-13.18/-46.7-/-19.31/-29.39** with normalization, showing the superiority of our method in safety-critical deployment.

Moving to compare energy and normalized energy functions, we show an improvement of **+1.82/+19.00/+4.18/+16.1** on the AUROC metric. Similar to the results on max logits, we observe a significant improvement over the Anomaly Track and Road Anomaly. Given that both datasets contain online-sourced images, we show the gain of normalizing in a high domain shift on image composition. On Average Precision, following the consistency of improvement, normalizing results in an increase of **+20.68/+43.71+/+41.32/+19.88**. Similar to max logits, the FP rate reduces by **-13.96/-50.12/-23.51/-30.41**.

Compared to the gain on max logits, normalizing the class score yields a higher increase in the accuracy of the energy score. This difference is consistent with different metrics and different backbones. This observation suggests that not only the max class score but all class scores are projected into a more robust representation with

Method	Lost and Found			Anomaly Track			Obstacle Track			Road Anomaly		
	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR
Norm - N	15.77	0.70	99.99	21.36	8.92	98.03	15.54	0.37	99.99	32.43	7.05	98.82
Max Softmax [24]	92.63	26.16	29.46	76.23	33.72	63.45	93.41	10.62	21.25	65.00	14.02	74.62
Max Logit [23]	97.24	63.37	16.54	77.51	34.52	65.79	95.81	41.04	22.08	71.10	17.76	71.36
SML [25]	92.39	50.71	50.39	62.82	19.05	76.55	80.47	18.76	70.34	57.23	11.38	89.45
Entropy [24]	93.82	43.96	28.74	77.52	36.90	63.44	94.80	25.57	20.56	66.51	15.46	74.37
Energy [32]	97.44	66.63	15.48	76.58	31.57	65.33	95.66	50.35	23.69	72.14	18.93	70.83
N-Max logit	98.90	78.18	3.26	94.22	69.22	19.09	99.20	68.14	2.77	82.84	26.56	41.97
N-Energy	**99.26**	87.31	**1.52**	95.58	75.28	17.18	**99.84**	91.67	**0.14**	**88.24**	38.41	**35.65**
N-Energy + CW	99.25	**89.70**	1.77	**96.76**	**83.13**	**15.12**	99.77	**93.78**	0.18	88.10	**42.87**	40.42

Table 5.2: **Effect of normalization on OOD measures using OCRNet[54].** We follow the same experimental structure as the DeepLabv3+[10] experiment.

hypersphere projection in terms of OOD detection.

Unlike the DeepLabv3+, we observe comparable results on energy function with and without increased class weight. However, there is a consistent gain on Average Precision by **+2.39/+7.85/+2.11/+4.46**. We optimize for class weight η using Anomaly Track. Consequently, on the Anomaly Track, class weights resulted in an improvement over all metrics. As we employ a coarse grid search on η, these results can be further improved with a denser search.

Observed improvements with normalization are empirical evidence for the ambiguity of class scores maintained by the norm of the embedding. Projecting all embeddings in a hypersphere representation with a constant norm renders class scores solely dependent on the whatness of an object, thus yielding a better OOD detection accuracy. Our results show that post-hoc decoupling is an employable strategy since it improves the OOD detection accuracy significantly.

Normalization layers of visual transformers implicitly yield an embedding space where norms are approximately equal. Using a visual transformer backbone is equivalent to using a backbone with a hyperspherical embedding space. To demonstrate this, we calculate the norms' mean and standard deviation with the Mask2Former employed with the Swin-L[35] backbone. We observe that on four datasets: Lost and Found, Anomaly Track, Obstacle Track, and Road Anomaly, the mean of the norms of the embedding is **13.90/13.89/13.89/13.89** for ID and **13.88/13.90/13.88/13.90** for OOD respectively, with a standard deviation of approximately **0.01** for all cases.

Method	Lost and Found			Anomaly Track			Obstacle Track			Road Anomaly		
	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR	AUC	AP	FPR
RBA-R101[37, 22]	69.16	7.21	98.97	80.60	60.93	95.49	68.75	24.14	99.87	73.71	38.19	87.59
EAM-R101[18, 22]	70.94	6.51	98.03	81.13	60.69	90.74	76.87	25.56	99.99	76.76	39.97	84.24
En-R101[32, 22]	71.76	7.26	86.46	81.39	61.14	90.14	82.68	26.54	99.99	78.23	39.72	83.47
RBA-SwinL[37, 35]	80.63	27.93	97.13	91.10	85.24	86.11	98.15	94.47	0.25	94.16	77.76	28.69
EAM-SwinL[18, 35]	81.94	28.22	96.10	94.17	87.05	33.11	99.34	95.05	0.21	94.49	**78.45**	22.49
En-SwinL[32, 35]	83.31	28.02	88.11	94.67	**87.31**	30.50	99.51	**95.13**	0.21	**94.60**	77.93	**22.49**
OCRNet-R101[54, 22]	**99.18**	**87.42**	**2.41**	**96.76**	83.13	**15.12**	**99.77**	93.78	**0.18**	88.10	42.87	40.42

Table 5.3: **Comparison of our method to recent unsupervised methods that utilize Mask2Former [11] architecture.** Our method outperforms the previous methods when utilized with the same backbone. It shows comparable results when compared backbones are utilized with a ViT [15, 35] backbone. Naming convention follows method-backbone. En is energy.

We claim that increased OOD detection accuracy using a visual transformer backbone in recent methods is mainly related to the normalization property of the attention layers. One backbone that yields a significantly higher OOD detection accuracy is Mask2Former[11]. OOD detection methods that use Mask2Former commonly refer to the architecture of masked attention layers as a reason for increased accuracy and build OOD detection methods based on said architecture[1, 37, 18]. To demonstrate the driving effect on OOD detection accuracy on Mask2Former, we propose an experiment comparing a simple energy score with two recent OOD detection methods on Mask2Former.

In Table 5.3, we evaluate two unsupervised OOD detection methods that claim to leverage the architecture of Mask2Former. RBA claims that a pixel rejected by all of the class masks of Mask2Former can be classified as OOD. Extending on this, EAM extracts an anomaly score for the whole image from the class token of Mask2Former. It distributes this anomaly score to per-mask features, resulting in a single mask of anomaly scores. We compare these methods with a simple energy score calculated from the class scores of the last layer of Mask2Former. We use the same Mask2Former models with R-101(ResNet-101) and Swin-L backbones we acquired from the official repository of Mask2Former to compare the three methods. For the implementation of RBA and EAM, we again use the original implementation from the repositories of said methods.

Comparison between R-101 and Swin-L results shows the significant effect of the backbone regardless of the model architecture. For instance, the change between AP of R-101 and Swin-L backbones in Obstacle Track for three methods with Mask2Former is **+68.59/+69.49/+70.33**. FPR dramatically reduces from nearly 100 to nearly 0 , just with the backbone change. This dramatic change is consistent with all datasets. To compare the effects of different methods, we emphasize the importance of building an experiment where all methods are tested with the same backbone. Moving to the impact of different OOD score extraction methods, we show that a simple energy function on the Swin-L backbone outperforms the other two methods in 10 of the 12 experiments. Further, our OCRNet method with norm normalization and R-101 backbone outperforms the Mask2Former methods with the same backbone in all experiments and stays competitive with their Swin-L counterpart in 7 out of 12 experiments. Our findings suggest that the high OOD detection accuracy of Mask2Former methods does not come from either Mask2Former architecture or OOD score methods but mainly from backbone and normalized embeddings as a side product of visual transformer backbones.

5.5 Qualitative Experiments

We visualize the output score maps using a heat scale in qualitative evaluation. On the heat scale, low values are represented by black, and high values are bright yellow. For visualization, we normalize the OOD scores using the formula:

$$s_n(x) = \frac{s(x) - min_i(s_i(x))}{max_i(s_i(x)) - min_i(s_i(x))} \tag{5.5}$$

where $s_n(x)$ is the normalized class score, brought into $0-1$ scale, $s(x)$ is the original score, and $min_i(s_i(x))$ and $max_i(s_i(x))$ are image-wise calculated minimum and maximum score. We normalize the class score image-wise and not globally for visualization purposes.

Figure 5.10 shows a comparison between energy and normalized energy. The top two images contain an animal on the scene as an OOD. We observe that this is a case of failure for the energy score, as OOD object have an overall low free energy. Also,

ID segments, such as the intersection of vegetation and the road in the first image and the side of the road in the second image, have relatively higher free energy. With normalized energy, not only do OOD pixels have higher energy scores, but the high energy of the ID regions is relatively normalized. This normalization can be observed in the intersection of ID segments in the first image and the side road in the second image. To further demonstrate this effect, we present a scene with a telephone box as OOD in the third image. Even though both methods have high energy for OOD pixel, the energy function, compared to the normalized energy, has more false negatives in the distant regions. In the fourth image, where the dog is OOD, we see high scores for the OOD pixels before normalization. However, similar to the third image, we observe a high false positive rate on the road. Again, this high energy on the ID is mitigated by score normalization. We see a fox as OOD in the fifth image and zebras as OOD in the sixth image. Similar to the first two images, OOD pixels have a low score before the normalization. This is also due to instance-oriented classes in Cityscapes, such as vehicles, having a high score when a distinct object is on the road. Normalization mitigates this effect and yields a high OOD score for the OOD pixels. The last image exemplifies a high domain shift from the noisy and blurred regions due to rain. As a result, the model outputs low scores and high energy throughout the image. The tire, as OOD, has a relatively low score compared to the rest of the image. Normalization mitigates high energy resulting from the domain shift, and the tire can be discriminated.

To illustrate the direct effect of normalization on class scores, we extend our qualitative evaluation to max logits. In Figure 5.11, we compare the heat maps of max logits with normalized max logits over multiple images. The first image shows multiple animals as OOD. Following our claims on the ambiguity of class scores, the max logit score on animals is on a similar scale to the rest of the image. Only the edge of the animals has a relatively high max logits score, which is expected since it is also observed in ID objects. With normalization, animals, as a whole, have a high max logit score, which discriminates them from the rest of the image. In the second image, a small toy on the road is OOD. Even though both models can detect the toy, normalization reduces the score of the ID pixels. In the third image, the max logit score partly detects the OOD tractor. We see areas of low energy, such as areas around tires. With normalized max logits, the tractor as a whole has a high energy

score. The fourth image shows an umbrella and two cups on the road as OOD. Even though there is no significant difference in the score of the OOD pixels between max logits and normalized max logits, OOD detection improves with a reduced FP rate. In the fifth image, pylons as OOD show a similar score distribution to the animals in the first image under max logits. Similar to the previous examples, by normalization, scores of the OOD pixels is increased, and ID pixels decreased. The same trend persists with the sixth and seventh images, where a bobby car and two boxes are OOD, respectively.

Qualitative evaluation shows us that with normalization, OOD detection accuracy not only improved by increased scores on OOD pixels but also by decreased scores on ID pixels. With max logits and the energy score, we observe that ID scores blend together to a low value. Coupled with the consistent increase in the OOD scores, this explains the significant improvement in the quantitative results.

5.6 Additional Experiments

Via additional experiments, we provide further insights into the effects of norm and class weights.

In the first experiment, we reintroduce the norm to the class scores and present the effects on max logits. For reintroducing, we increase an exponent over norms ranging between 0 to 1. By recalling the Formula 4.17, we formalize the reintroducing as follows:

$$s'_e = \|z\|^{\delta} s' \tag{5.6}$$

where s'_e is experimental class score, and δ is increased in the said range. $\delta = 1$ is equivalent to normalizing the class scores by norms, and $\delta = 0$ results in the original class scores. Following the formulation, max logits on experimental scores become:

$$S'_e(x) = -max_c \frac{f(w_c, x)}{\|z\|^{1-\delta}} \tag{5.7}$$

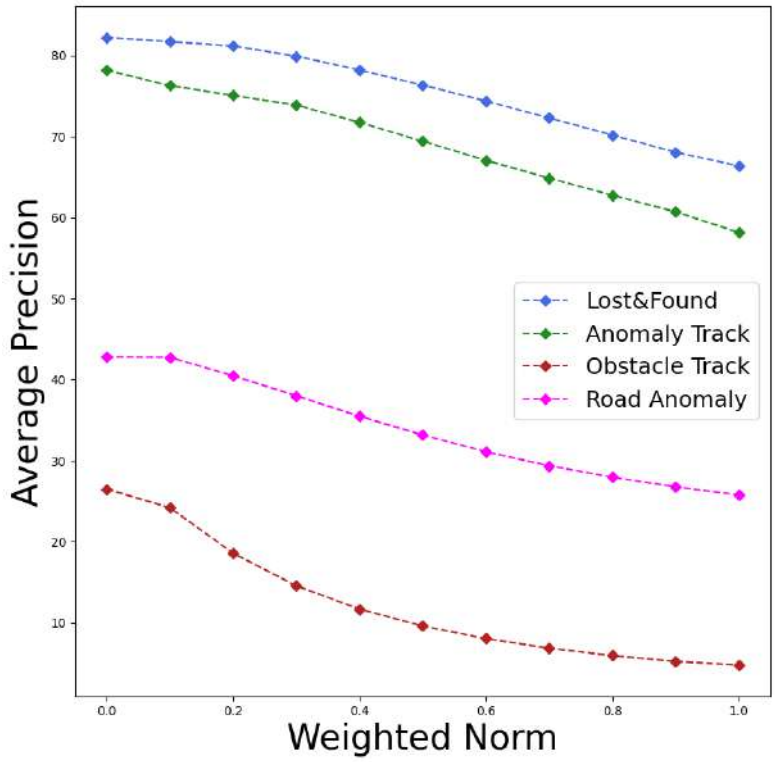

Figure 5.5: **The effect of δ on Average Precision using DeepLabv3+ [10].** The x-axis contains the increasing values of δ. We present the impact of reintroducing the norm on Average Precision with four datasets. We observe a consistent decrease with increasing δ.

Figure 5.5 shows the change of Average Precision with increasing δ using DeepLabv3+. Dots in the graph show sample points. Our experiments show the applicability of a weighted sum on norms and normalized scores in the segmentation domain. We observe that reintroducing norms, even with small weights, results in a decrease in the Average Precision metric. In Figure 5.6, we show the results of the same experiment with the OCRNet. Similar to the DeepLabv3+, we observe a consistent decrease in the accuracy with the increased weight of δ. Our empirical findings in this experiment oppose a weighted sum approach [56].

In our second experiment, we investigate the effects of the exponent η over class weights. We use Anomaly Track as a reference to optimize for η. Figure 5.7 shows the effect of changing η on Average Precision. We show that, as suggested by previous work for the classification domain, normalizing by class weights is not adaptable to the segmentation domain. $\eta = 0$ corresponds to normalizing, which results in a

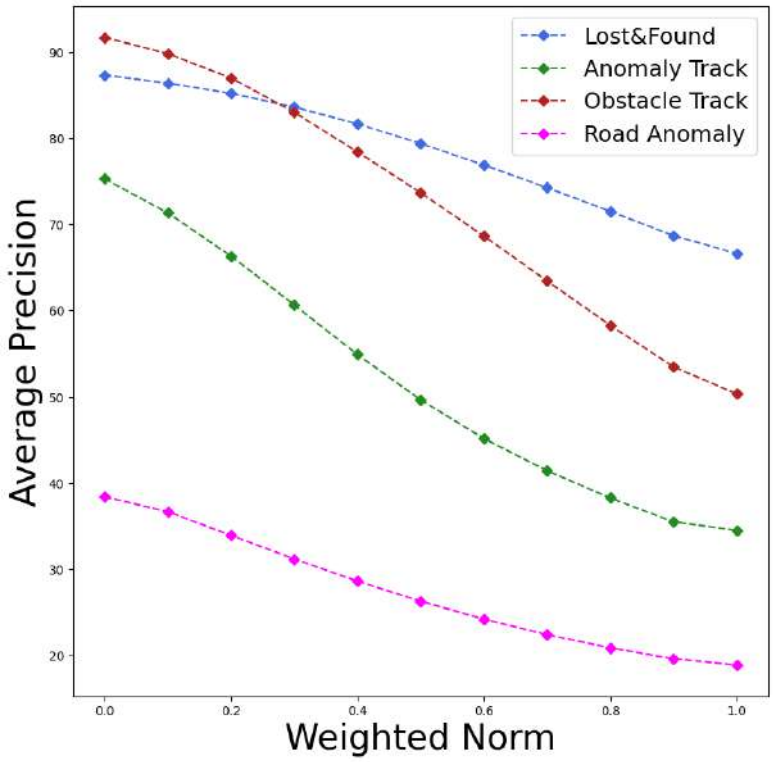

Figure 5.6: **The effect of δ on Average Precision using OCRNet [54].** Using the same experimental structure, we observe a higher impact on the norm compared to DeebLabv3+ [15]. Consistency of the effect on Average Precision also persists with OCRNet.

significant decrease over the original class scores. We see an increasing trend with increasing η. However, gain diminishes after $\eta = 1.5$ for both of the backbones. After the gain diminished, we observed that FP rates started to increase slightly, and as a result, we selected $\eta = 1.5$ for the rest of our experiments.

5.7 Limitations

In Section 4.2, we discussed the assumption of similar external distribution between ID and OOD objects. Our claim implies that norms can be a discriminative measure if there is a distinct difference in external distribution between ID and OOD. In that case, removing the norms could result in a loss of accuracy because norm decoupling would remove a discriminative channel from the class scores.

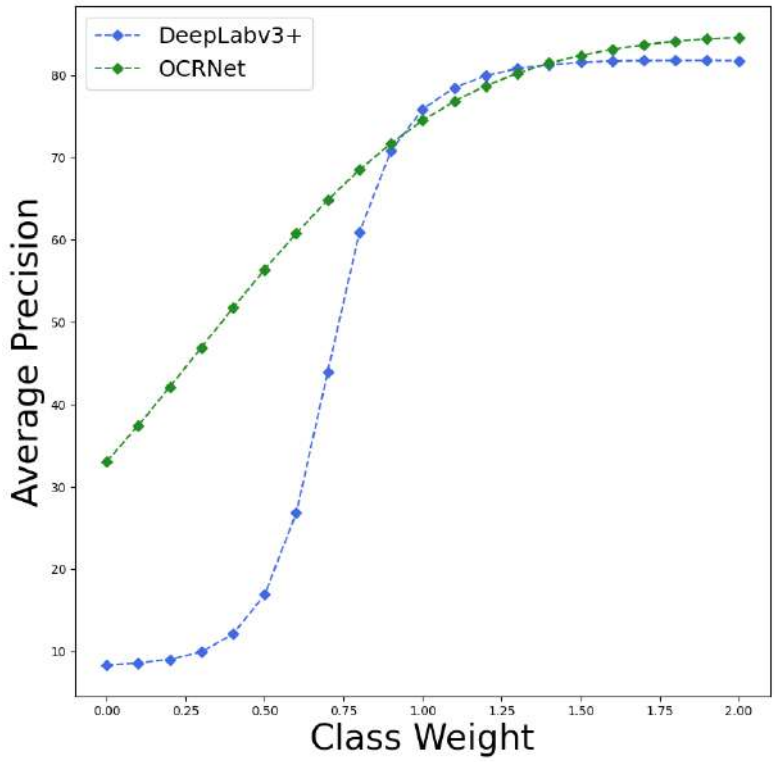

Figure 5.7: **Average Precision under varying η values.** We show the findings for optimizing η with Anomaly Track, which we choose as a reference set on optimization. We present results with DeepLabv3+ [10] and OCRNet [54] architectures.

To evaluate this case, we experiment with a dataset with an inherent difference in external distributions between ID and OOD. In Figure 5.8, we show example images from the FS Static [4] dataset. FS Static contains images from the Cityscapes validation set with pasted OOD images from online sources. While pasting OOD patches, blending is applied to the images, reducing the OOD object's transparency. Since OOD images are from external sources and have a different transparency compared to the rest of the images, it creates a scenario where our assumption of similar external distributions does not hold.

Figure 5.9 shows the distribution of norms of embeddings sampled from the FS Static dataset. Recalling our experiment in Figure 4.5, we show that the difference of the norms with FS Static is parallel to the shift resulting from Gaussian blur and changed color density. This parallel shift confirms our claims on both the FS Static dataset's external distribution and the external shift's effect on the norms.

Figure 5.8: **Example Images on FS Static [4].** Presents a challenge in a domain where our assumption of external distribution does not hold.

Method	DeepLabv3+			OCRNet		
	AUC	AP	FPR	AUC	AP	FPR
Max Logit [23]	88.87	22.72	**65.10**	94.90	27.75	20.51
Energy [32]	**88.99**	**23.82**	65.13	**95.05**	29.30	**20.30**
N-Max Logit	85.34	19.99	73.71	93.44	**30.00**	27.34
N-Energy	85.14	15.49	70.43	93.20	29.19	28.11

Table 5.4: **Quantitative results of normalization with FS Static dataset [4].** We present results with DeepLabv3+ [10] and OCRNet [54] architectures.

Table 5.4 shows the quantitative evaluation of normalized class scores on FS Static. We show results with DeepLabv3+ and OCRNet following our previous experimental setup. With DeepLabv3+ and OCRNet, we observe a decrease in OOD detection accuracy when normalized class scores are employed with max logits and energy. With OCRNet, even though there is no significant change in Average Precision, we observe a decrease in AUROC and an increase in FP rate. Even though this experiment shows a limitation for our method, it also is empirical evidence for our motivation and technical explanation of the effects of the norm on class scores.

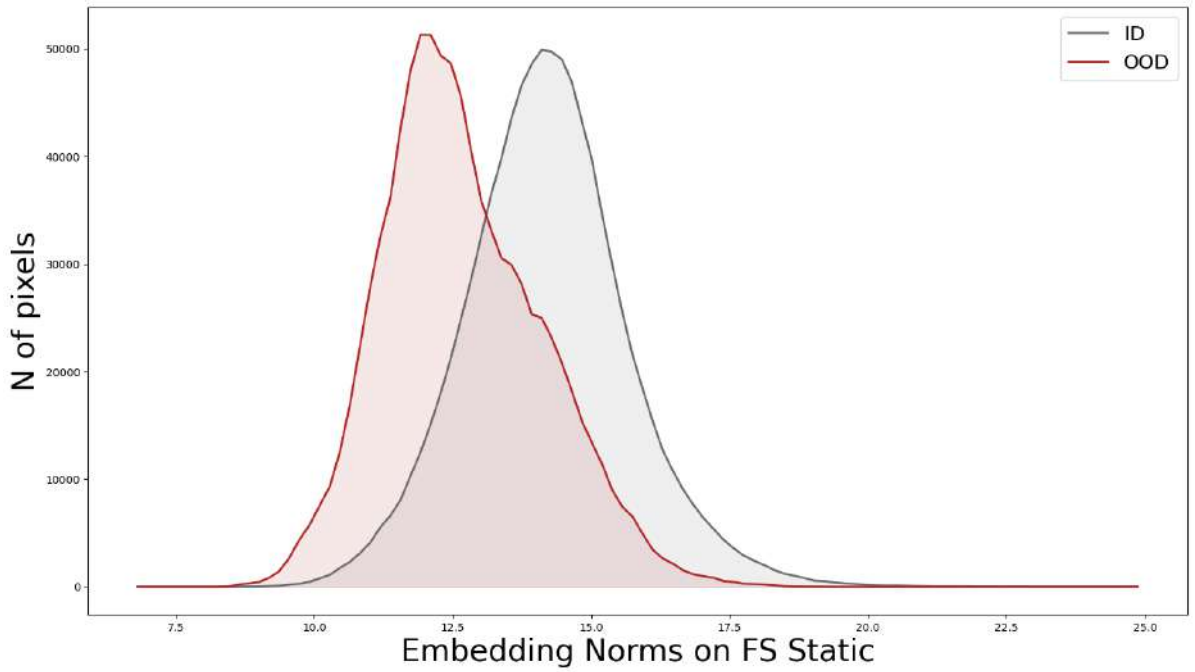

Figure 5.9: **Distribution of the norms of the embeddings from FS Static dataset [4].** We observe a similar shift to our external distribution experiment in Figure 4.5.

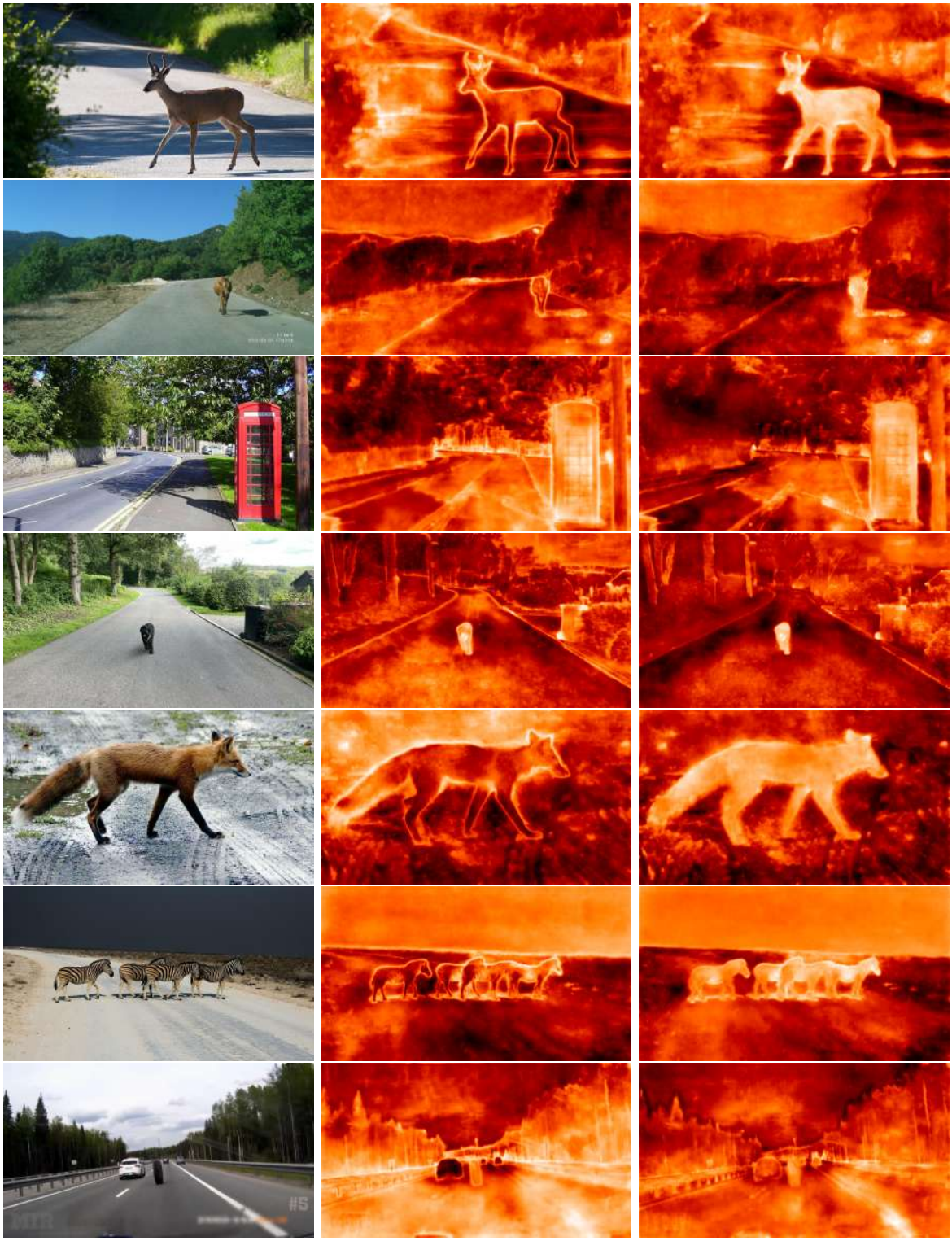

Figure 5.10: **Qualitative comparison of energy [32] and normalized energy on OCR-Net [54] .** We show the original image, energy, and normalized energy from left to right.

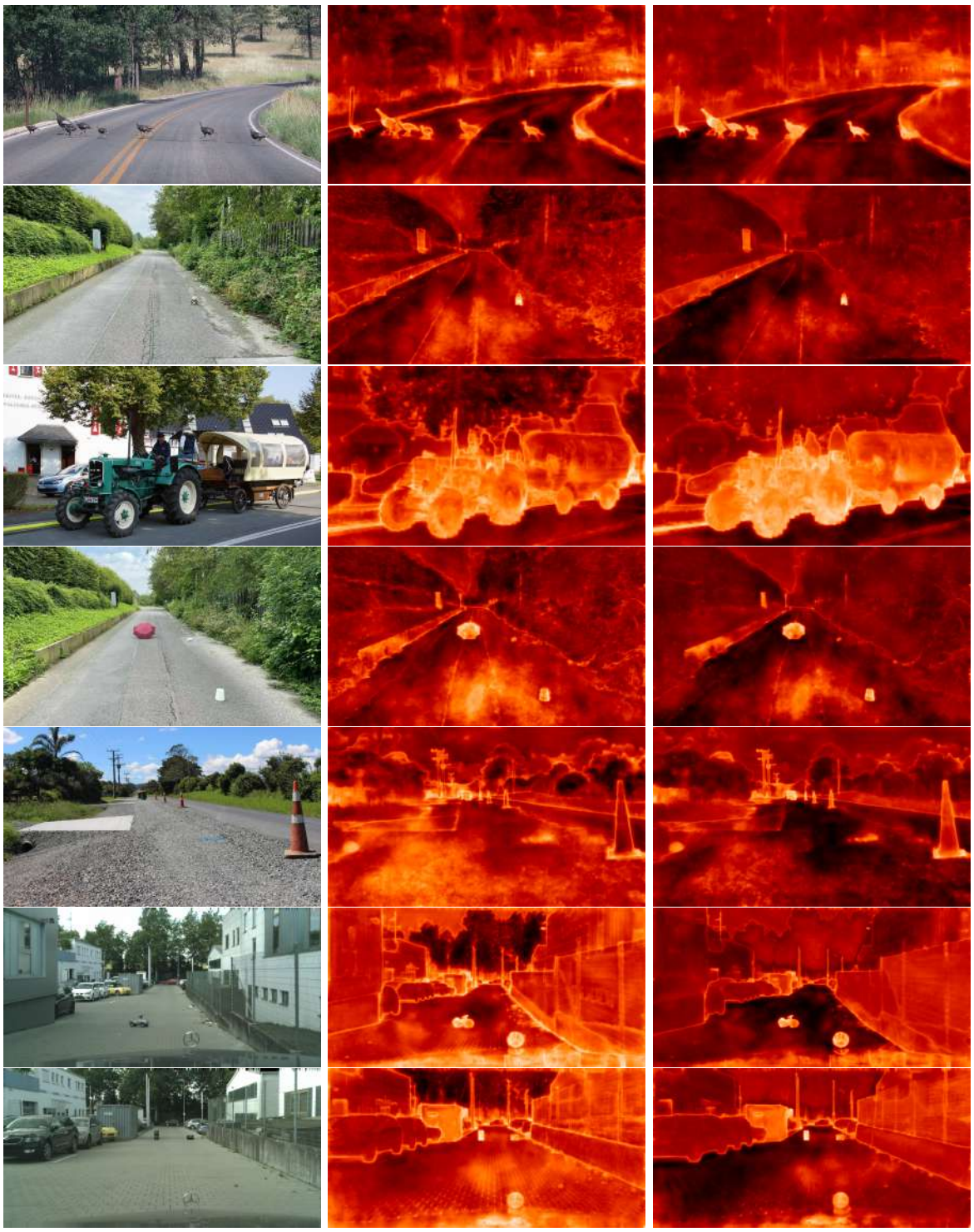

Figure 5.11: **Qualitative comparison of max logits [23] and normalized max logits on OCRNet [54].** We show the original image, max logits, and normalized max logits from left to right.

6 Conclusion

This work extends the post-hoc norm decoupling for OOD detection to the segmentation domain. Through technical investigation, we reveal the ambiguity of embedding norms and propose a novel explanation of their effects on class scores. Backed by empirical evidence, we provide insights about how this ambiguity manifests itself. Following our explanation, we present a solution for the ambiguity of class scores imposed by the norms. By projecting embeddings to a hypersphere, we ensure a representation that is robust to external factors other than the object's whatness. We demonstrate the utilization of normalized class scores as an OOD measure by employing them with max logits and energy functions. Our empirical findings suggest that class weight normalization for decoupling in classification does not extend to the dense segmentation domain. On the contrary, increasing class weights further diverges ID and OOD distributions, improving accuracy.

We back our claims by evaluating our method on multiple datasets with two different backbones. Our method consistently improve score-based OOD detection methods by a significant margin. This consistency is empirical evidence for our technical findings. Further, we reason the improved OOD detection accuracy yielded by visual transformer backbones. We show the practical equivalence of hyperspherical and ViT [15] representations and achieve comparable results to recent ViT models by normalizing class scores. Our empirical findings highlight the importance of the backbone in OOD detection.

With this work, we emphasize the importance of understanding the model's decision mechanisms in OOD detection. We hope our insights can inspire further development in OOD detection and explainability domains.

List of Figures

List of Tables

Bibliography

[1] J. Ackermann, C. Sakaridis, and F. Yu. "Maskomaly: Zero-Shot Mask Anomaly Segmentation." In: *BMCV*. 2023.

[2] G. D. Biase, H. Blum, R. Siegwart, and C. Cadena. "Pixel-wise Anomaly Detection in Complex Driving Scenes." In: *CVPR*. 2021.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[4] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena. "The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation." In: *IJCV* (2021).

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. "Emerging Properties in Self-Supervised Vision Transformers." In: *ICCV*. 2021.

[6] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann. "SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation." In: *NIPS*. 2021.

[7] R. Chan, M. Rottmann, and H. Gottschalk. "Entropy Maximization and Meta Classification for Out-Of-Distribution Detection in Semantic Segmentation." In: *CVPR*. 2021.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." In: *TPAMI*. 2017.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking Atrous Convolution for Semantic Image Segmentation." In: *arXiv e-prints* (2017), arXiv–1706.

[10] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." In: *ECCV*. 2018.

[11] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. "Masked-attention Mask Transformer for Universal Image Segmentation." In: *CVPR*. 2022.

[12] B. Cheng, A. G. Schwing, and A. Kirillov. "Per-Pixel Classification is Not All You Need for Semantic Segmentation." In: *NIPS*. 2021.

[13] J. H. Cho, U. Mall, K. Bala, and B. Hariharan. "PiCIE: Unsupervised Semantic Segmentation Using Invariance and Equivariance in Clustering." In: *CVPR*. 2021.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. "The cityscapes dataset for semantic urban scene understanding." In: *CVPR*. 2016.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." In: *ICLR*. 2020.

[16] X. Du, X. Wang, G. Gozum, and Y. Li. "Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild." In: *CVPR*. 2022.

[17] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. "Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One." In: *arXiv e-prints* (2020), arXiv–1912.

[18] M. Grcic, J. Šarić, and S. Šegvić. "On Advantages of Mask-level Recognition for Outlier-aware Segmentation." In: *CVPR Workshop*. 2023.

[19] M. Grcić, P. Bevandić, Z. Kalafatić, and S. Šegvić. "Dense Out-of-Distribution Detection by Robust Learning on Synthetic Negative Data." In: *CoRR*. 2018.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks." In: *CVPR*. 2016.

[21] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman. "Unsupervised Semantic Segmentation by Distilling Feature Correspondences." In: *ICLR*. 2022.

[22] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *CVPR*. 2016.

[23] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song. "Scaling Out-of-Distribution Detection for Real-World Settings." In: *ICML*. 2022.

[24] D. Hendrycks and K. Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks." In: *ICLR*. 2017.

[25] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo. "Standardized Max Logits: A Simple yet Effective Approach for Identifying Unexpected Road Obstacles in Urban-Scene Segmentation." In: *ICCV*. 2021.

[26] Y. Lecun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. "A tutorial on energy-based learning." In: *Predicting structured data*. 2006.

[27] K. Lee, K. Lee, H. Lee, and J. Shin. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." In: *NIPS*. 2018.

[28] C. Liang, W. Wang, J. Miao, and Y. Yang. "GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models." In: *NIPS*. 2022.

[29] S. Liang, Y. Li, and R. Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks." In: *ICLR*. 2018.

[30] K. Lis, S. Honari, P. Fua, and M. Salzmann. "Detecting Road Obstacles by Erasing Them." In: *CoRR*. 2023.

[31] K. Lis, K. Nakka, P. Fua, and M. Salzmann. "Detecting the Unexpected via Image Resynthesis." In: *ICCV*. 2018.

[32] W. Liu, X. Wang, J. D. Owens, and Y. Li. "Energy-based Out-of-distribution Detection." In: *NIPS*. 2020.

[33] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro. "Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation." In: *ICCV*. 2023.

[34] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. "Swin Transformer V2: Scaling Up Capacity and Resolution." In: *CVPR*. 2022.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." In: *ICCV*. 2021.

[36] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. "Image Segmentation Using Deep Learning: A Survey." In: *TPAMI*. 2022.

[37] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney. "RbA: Segmenting Unknown Regions Rejected by All." In: *ICCV*. 2023.

[38] guyen Ngoc-Hieu, N. Hung-Quang, T.-A. Ta, T. Nguyen-Tang, K. D. Doan, and H. Thanh-Tung. "A Cosine Similarity-based Method for Out-of-Distribution Detection." In: *ICML Workshop*. 2023.

[39] Y. Ohta, T. Kanade, and T. Sakai. "An Analysis System for Scenes Containing objects with Substructures." In: *IJCPR*. 1978.

[40] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. "DINOv2: Learning Robust Visual Features without Supervision." In: *arXiv e-prints* (2023), arXiv–2304.

[41] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. "Deep Learning for Anomaly Detection: A Review." In: *ACM Computing Surveys* (2021).

[42] J. Park, J. C. L. Chai, J. Yoon, and A. B. J. Teoh. "Understanding the Feature Norm for Out-of-Distribution Detection." In: *ICCV*. 2023.

[43] J. Park, Y. G. Jung, and A. B. J. Teoh. "Nearest Neighbor Guidance for Out-of-Distribution Detection." In: *ICCV*. 2023.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." In: *ICML*. 2021.

[45] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *MICCAI*. 2015.

[46] C. E. Shannon. "A mathematical theory of communication." In: *The Bell System Technical Journal* (1948).

[47] E. Shelhamer, J. Long, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation." In: *PAMI*. 2017.

[48] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro. "Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes." In: *arXiv e-prints* (2021), arXiv–2111.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is All you Need." In: *NIPS*. 2017.

[50] S. Weber, B. Zöngür, N. Araslanov, and D. Cremers. "Flattening the Parent Bias: Hierarchical Semantic Segmentation in the Poincaré Ball." In: *arXiv e-prints* (2024), arXiv–2404.

[51] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li. "Mitigating Neural Network Overconfidence with Logit Normalization." In: *ICML*. 2022.

[52] F. Yu and V. Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions." In: *ICLR*. 2016.

[53] Y. Yu, S. Shin, S. Lee, C. Jun, and K. Lee. "Block Selection Method for Using Feature Norm in Out-of-Distribution Detection." In: *CVPR*. 2023.

[54] Y. Yuan, X. Chen, and J. Wang. "Object-contextual representations for semantic segmentation." In: *ECCV*. 2020.

[55] R. Zhang, P. Isola, and A. A. Efros. "Colorful Image Colorization." In: *ECCV*. 2016.

[56] Z. Zhang and X. Xiang. "Decoupling MaxLogit for Out-of-Distribution Detection." In: *CVPR*. 2023.