More Than Accuracy: An Empirical Study of Consistency Between Performance and Interpretability

Yun Du[1], Dong Liang[1](✉), Rong Quan[1], Songlin Du[2], and Yaping Yan[2]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
{muyun,liangdong,quanrong21}@nuaa.edu.cn

[2] Southeast University, Nanjing 210096, China
{sdu,yan}@seu.edu.cn

Abstract. Expected calibration error (ECE) is a popular metric to measure and calibrate the inconsistency between the classification performance and the probabilistic class confidence. However, ECE is inadequate to reveal why the deep model makes inconsistent predictions in specific samples. On the other hand, the class activation maps (CAMs) provide visual interpretability, highlighting focused regions of network attention. We discover that the quality of CAMs is also inconsistent with the model's final performance. In this paper, to further analyze this phenomenon, we propose a novel metric—VICE (Visual Consistency), to measure the consistency between performance and visual interpretability. Through extensive experiments with ECE and VICE, we disclose that the model architectures, the pre-training schemes, and the regularization manners influence VICE. These phenomena deserve our attention, and the community should focus more on a better trade-off in model performance and interpretability.

Keywords: Visual consistency · Expected calibration error · Model interpretability

1 Introduction

Besides pursuing superior performance via deep neural networks, its interpretability has increased attention in risk-sensitive real-world scenarios. In scenarios such as finance and medical care, where industry users are eager to know how the AI systems make decisions, deep neural networks still have a long way to yield complete interpretability. Trustworthy AI is less likely to be built by black-box models. It is thus a significant challenge to achieve an effective balance between performance and interpretability.

Expected calibration error (ECE) [22] is a metric to measure the inconsistency between the final classification performance and the probabilistic class confidence. Minimizing ECE could yield more accurate calibrated confidence.

S. Khanna et al. (Eds.): PRICAI 2022, LNCS 13631, pp. 579–590, 2022.
https://doi.org/10.1007/978-3-031-20868-3_43

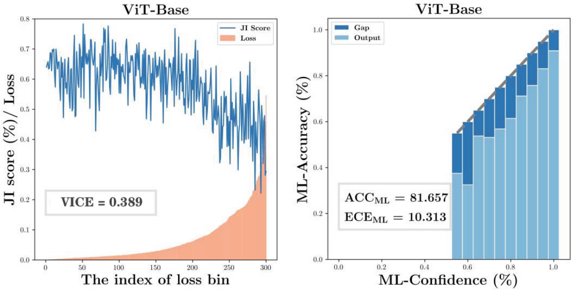

Fig. 1. The reliability diagrams on VOC2012Aug.

However, ECE and the calibrated confidence are inadequate to reveal why the deep model makes inconsistent predictions. As shown in the bottom right of Fig. 1, from the reliability diagrams of ECE, we can observe that the deep model is prone to over-confidence.

On the other hand, the class activation maps (CAMs) [34] is widely used to provide model's interpretable cues, providing consistent visual interpretability with human cognition. As shown in the bottom left of Fig. 1, the quality of CAMs is inconsistent with the sample loss. The average quality of the CAMs varied considerably between samples in neighbouring loss bins.

In this paper, we propose a novel metric—VICE, to measure the consistency between performance and visual interpretability. The top figure of Fig. 1 compares the differences between ECE and VICE, which are defined explicitly in Sect. 2 and Sect. 3, respectively. Our contribution can be summarized as follows,

(1) We propose VICE as a novel sample-wise fine-grained metric, to evaluate the consistency between model performance and visual interpretability in a new perspective.
(2) With case studies on five heterogeneous models, we found a series of inspiring conclusions involving the model capacity and architectures, the pre-training schemes, and the regularization mechanisms. Specifically, blindly increasing the model capacity may harm model interpretability; models with visual attention have better visual interpretability but worse consistency between model performance and visual interpretability; self-supervised pre-training and additional regularization learn more interpretable feature representation. These guide us to leverage more significant impact factors for trustworthy model design.

2 The ECE Metric

2.1 Revisiting the Expected Calibration Error

Expected calibration error (ECE) [22] is a popular tool to measure the distribution difference between the probabilistic confidence and the classification accuracy. More detailed, the samples are evenly partitioned into M bins according to their confidence scores, measuring the differences between accuracy and confidence. ECE allows us to better judge the risk of the model's unexpected predictions.

2.2 Extend ECE for Multi-label Classification

Here, we first extend ECE to multi-label (ML) classification, which is a better approximation of realistic scenarios and more approaching to human cognition than single-label classification. Since multi-label classification assigns multiple labels to each sample, we define ML-accuracy (Eq. 1) and ML-confidence (Eq. 3) of samples in each bin, to enable ECE to realize evaluation under multi-label classification.

$$\mathrm{acc_{ML}}(\mathrm{B_m}) = \frac{1}{|B_m|}\, sum_{i \in B_m} \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \tag{1}$$

In Eq. 1, y_i is the true set of labels, $\hat{y}_i$ be the predicted set of labels. ML-accuracy is measured symmetrically how close y_i is to $\hat{y}_i$ [9].

$$\mathrm{conf_{ML}^{i}} = \begin{cases} \frac{1}{C}\sum_{c=1}^{C} \hat{p}_{i,c} \cdot \mathbf{1}(\hat{p}_{i,c} \geq t) \ , \exists\ \hat{p}_{i,c} \geq t \\ max(\hat{p}_{i,c}) \ , \forall\ \hat{p}_{i,c} < t \end{cases} \tag{2}$$

$$\mathrm{conf_{ML}}(\mathrm{B_m}) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathrm{conf_{ML}^{i}} \tag{3}$$

Equation 3 calculates the ML-confidence. If the confidence of a category $\hat{p}_{i,c}$ is greater than a threshold t (usually be set as 0.5), the model believes that the category is present in the image. So the confidence for that sample is represented by the average of all the confidence above this threshold. If the model does not exceed the threshold for all categories, then the maximum posterior probability is taken as the confidence of the sample.

$$\mathrm{ECE_{ML}} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc_{ML}}\left(B_m\right) - \mathrm{conf_{ML}}\left(B_m\right)\right| \tag{4}$$

Equation 4 extends ECE to $\mathrm{ECE_{ML}}$ for multi-label classification.

3 The Proposed VICE Metric

3.1 Motivation

In many trials, we found an inconsistency between the performance of the deep learning models and their visual interpretability. Some samples with low loss have poor visual interpretability, with incomplete CAM coverage, while some other samples with relatively complete CAMs have biased predictions, as shown in Fig. 1. Such inconsistencies lead to restricted and unreliable applications. A model with good interpretability is not only consistent in confidence and accuracy (with lower ECE). In applications with a firm reliance on visual interpretability, we expect the good-performed models to be more interpretable and consistent with human perception. So we tend to introduce a sample-wise perspective to fine-grained measure the consistency of performance and visual interpretability, which is referred to as VICE (VIsual ConsistEncy). If a model has a good visual consistency, the visual interpretability output by the model will focus on the object itself for correctly predicted samples, and for incorrectly predicted samples we can find the reasons for poor performance in visual interpretability, such as label noise, background effects, etc.

In the following part of this section, we describe how to calculate the consistency of performance with visual interpretability. We first present two types of visual interpretability corresponding to CNN and Transformer architectures, then use a sample-wise loss to characterize the network performance on the corresponding samples, and finally calculate VICE to indicate the consistency.

3.2 The Visual Interpretability of CNNs and Transformers

Due to the different architectures of CNN and Transformer, the manner of generating CAMs is different. As for a CNN, we follow the setting from [34] to generate a CAM M^c_{cnn} for class c, which is calculated as:

$$M^c_{cnn} = \sum_k w^c_k f_k \tag{5}$$

M^c_{cnn} directly correlates with the importance of a particular spatial location for a specific class c and thus functions as visual interpretability of the category predicted by the network.

The approach in [2] is adopted to generate CAMs for vision transformer. The layer's operation on two tensors X and Y is denoted by $L^{(n)}(X, Y)$. The two tensors are the input feature map and weights for layer n. Relevance propagation follows the generic Deep Taylor Decomposition [21].

3.3 The Quality for Visual Interpretability

The segmentation labels of CAMs are related to human cognition, and the quality of a CAM is the degree to which it conforms to human cognition. To evaluate the quality of visual interpretability (CAMs), we use the mean Jaccard index to measure the consistency of CAMs and segmentation labels. The definition is shown in Eq. 6.

$$JI(M_{cam}, M_{gt}) = \frac{1}{C} \sum_{i=0}^{C} \frac{|M_{cam} \cap M_{gt}|}{|M_{cam} \cup M_{gt}|} \tag{6}$$

A higher JI score means the model has higher quality CAMs with better visual interpretability.

3.4 The Sample-wise Performance

The sample-wise multi-label loss is used as an indicator of the model performance,

$$
\begin{aligned}
loss_{sw}(x, y) = -\frac{1}{C}\sum_{c} y_c \log[(1 + \exp(-x_c))^{-1}] \\
+ (1 - y_c)\log[\frac{\exp(-x_c)}{1 + \exp(-x_c)}]
\end{aligned}
\tag{7}
$$

As shown in Eq. 7, the error is measured by calculating the sigmoid cross-entropy between the output layer and the labels. Equivalently, the binary cross-entropy loss is calculated for each category.

3.5 The Final VICE

We quantitatively estimate the consistency between performance and interpretability by measuring the pearson correlation coefficient between sample-wise loss and the quality of CAMs. And the '−' of $loss_{sw}$ is to allow positive correlation.

$$
\text{VICE} = \rho(loss, JI) = \frac{\sum_{i=1}^{n}\left(loss_i - \overline{loss}\right)\left(JI_i - \overline{JI}\right)}{\sqrt{\sum_{i=1}^{n}\left(loss_i - \overline{loss}\right)^2}\sqrt{\sum_{i=1}^{n}\left(JI_i - \overline{JI}\right)^2}}
\tag{8}
$$

For samples with low loss, the model with a higher VICE score can give better visual interpretability than with a lower score. While for samples with high loss, we can also observe the reasons for their prediction errors via their CAMs. As shown in the top figure of Fig. 1, the VICE and ECE_{ML} evaluate the consistency between performance and interpretability from different perspectives.

4 Empirical Evaluation

4.1 Experimental Setup

Our case study followed the standard multi-label classification setting. The metrics, model families, and datasets used are introduced next.

Metrics for model performance: mean average precision (mAP) over all categories and accuracy for multi-label (described in Eq. 1) is adopted for evaluating the performance of models. The larger the ACC and mAP, the better the model performance.

Metrics for the consistency between the performance and interpretability: ECE_{ML} (described in Eq. 4) with 20 bins is to evaluate the consistency between accuracy and confidence. VICE (described in Eq. 8) is to measure the consistency between performance and visual interpretability. The smaller the ECE_{ML} and VICE, the greater the consistency of model performance and interpretability.

Model families. We select recent and historic state-of-the-art classification models, including ResNet [14], Res2Net [8], EfficientNet [26], ResNeSt [31], ViT [7]. If not explicitly mentioned, all of them are pre-trained on ImageNet1k [5] with full supervision.

Datasets. We evaluate on PASCAL VOC 2012 dataset with 21 class annotations. The official dataset separation has 1464 images for training, 1449 for validation, and 1456 for testing. Following the common experimental protocol, we take additional annotations from SBD [12] to build an augmented training set with 10582 images, which is named as VOC2012Aug. It provides labels for semantic segmentation, which allows us to evaluate the quality of the CAMs (Eq. 6). All metrics reported were performed on the validation set.

4.2 Models with Different Capacity

Table 1. Models with Different Capacity. The VICE does not monotonically increase with increasing model capacity.

Model	Params	mAP (%)↑/Acc (%) ↑	$\mathrm{ECE}_{\mathrm{ML}}$ (%)↓	VICE ↑
ResNet-18	11.177 M	83.917/72.188	14.897	0.185
ResNet-34	21.285 M	86.296/76.945	14.257	**0.226**
ResNet-50	23.508 M	87.470/78.617	**13.401**	0.215
ResNeSt-14	8.563 M	85.382/75.329	13.989	0.342
ResNeSt-26	15.020 M	87.395/79.170	13.248	0.311
ResNeSt-50	25.434 M	88.031/79.117	12.978	**0.322**
ResNeSt-101	46.226 M	88.856/80.415	**12.735**	0.272
ViT-Small	21.975 M	90.253/81.172	10.361	**0.393**
ViT-Base	86.416 M	91.238/81.657	**10.313**	0.389

Table 1 reports the model performance, $\mathrm{ECE}_{\mathrm{ML}}$, and VICE of the models with different capacities. As the model capacity increases, the mAP is increasing, and its $\mathrm{ECE}_{\mathrm{ML}}$ is decreasing, which is beneficial for practical tasks. But the VICE does not monotonically increase with increasing model capacity, which is confirmed in three architectures.

Discussion. Model capacity and representation capability have always been considered key to improving performance, but models are often over-parameterized. Increasing the model capacity would have a performance bottleneck and also increase the risk of overfitting. *Blindly increasing model capacity may harm visual interpretability and consistency*, which is unfavorable for scenarios highly dependent on visual interpretability.

4.3 Models with Various Architecture

To compare the impact of different model architectures fairly, we selected five models with similar model size and various feature extraction preferences. Figure 2 shows the $\mathrm{ECE}_{\mathrm{ML}}$ metrics for different models. From Eq. 3, if the confidence level of all classes is less than t, then the accuracy of this sample is equal

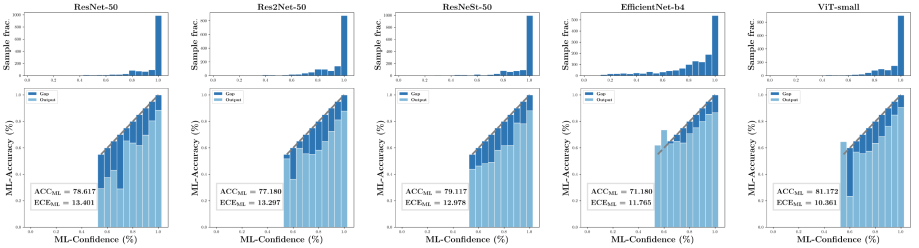

Fig. 2. We divided the samples into different bins according to the loss. The figure above shows the number of samples in each bin, and the figure below shows the accuracy and confidence in each bin. Regardless of the number of samples in each bin, the output prediction probabilities (light blue blocks) generally fall below the straight line y=x (dark blue blocks), which means that the deep model is prone to over-confidence. (Color figure oline)

Table 2. Models with Various Architecture. ResNeSt and ViT outperform ResNet in JI score, and their VICE fail to be better than that of ResNet.

Model	Params	mAP (%)↑/Acc (%) ↑	ECE_{ML} (%)↓	VICE ↑
ResNet-50	23.508 M	87.470/78.617	13.401	0.215
Res2Net-50	23.238 M	86.759/77.180	13.297	0.114
ResNeSt-50	25.434 M	88.031/79.117	12.978	0.322
EfficientNet-b4	17.549 M	83.690/71.180	11.765	**0.449**
ViT-Small	21.975 M	90.253/81.172	**10.361**	0.393

to zero, which is the reason why the accuracy in the bins with confidence below 0.5 is zero. ECE_{ML} is the calculation of the average discrepancy between the two statistical histograms. We find that the lower histogram shows that the output prediction probabilities (light blue parts) generally fall below the straight line $y = x$ (dark blue parts), which means that the deep model is prone to over-confidence. Moreover, according to Fig. 2, we can also conclude that the model calibration of the vision transformer is better than that of CNN. This inspires us to consider the potential of non-CNN models such as vision transformer, which not only unifies feature encoding architectures for multimodal tasks but also has better model calibration with lower ECE_{ML}. On the other hand, Table 2 shows the VICE score for different model architectures. ResNeSt and ViT outperform ResNet in terms of VICE score with comparable model sizes, suggesting that the CAMs of the models with integrated attention mechanisms are more complete. And EfficientNet-b4 has the highest VICE score, which hints at the potential of NAS-based architectures to achieve a better tradeoff between interpretability and performance.

Discussion. Models of different architectures have different consistency preferences, and the integration of multiple model architectures may improve interpretability for a given task.

4.4 Consistency Study with Self-supervised Pre-trained Models

Motivation. Recent approaches such as [1,3,4,10,13,18,28] show that the features extracted from networks training with a self-supervised learning paradigm can achieve better performance on downstream tasks that require more object semantics, such as object detection and semantic segmentation, which sparked the interest of many researchers. A large number of self-supervised learning algorithms focus on instance discrimination, considering a single image as a class, encouraging the model to bring the sample and its augmented image as close as possible on the feature space. MoCo [13] and SimCLR [3] with contrastive self-supervised learning even surpass the results of supervised learning. Reference [4] is the first work to apply self-supervision to Vision Transformer. DINO [1] explores the self-supervised approach so that ViT contains features related to semantic segmentation of images, which is useful for downstream tasks. Self-supervised models are more interpretable if they can explore target locations. Since the self-supervised learning paradigm is not constrained by category information, which allows the network to freely explore its suitable region of interest, rather than just focusing on the features about classification tasks. Therefore, we evaluate the performance of ResNet and ViT using different self-supervised schemes.

Table 3. Various Pre-train Setting with ResNet-50.

Model	Pre-train Dataset	Pre-train Alg	Finetuning	mAP (%) $\uparrow$/Acc (%) $\uparrow$	ECE_{ML} (%) $\downarrow$	VICE $\uparrow$
ResNet-50	ImageNet	Supervised	–	87.470/78.617	**13.401**	0.215
ResNet-50	ImageNet	BYOL	–	74.827/62.762	19.087	0.305
ResNet-50	ImageNet	DINO	–	73.147/57.978	18.080	0.203
ResNet-50	ImageNet	MoCov3	–	76.200/64.114	18.212	0.324
ResNet-50	VOC2012Aug	MoCov3	–	39.701/21.731	28.483	**0.348**
ResNet-50	ImageNet	Supervised	*fc*	84.492/74.028	**12.510**	0.260
ResNet-50	ImageNet	BYOL	*fc*	59.351/27.029	24.452	0.369
ResNet-50	ImageNet	DINO	*fc*	71.158/42.312	17.652	0.282
ResNet-50	ImageNet	MoCov3	*fc*	70.336/28.532	23.176	0.251
ResNet-50	VOC2012Aug	MoCov3	*fc*	38.105/19.587	29.203	**0.401**

Table 4. Various Pre-train Setting with ViT-Base.

Model	Pre-train Dataset	Pre-train Alg	Finetuning	mAP (%) $\uparrow$/Acc (%) $\uparrow$	ECE_{ML} (%) $\downarrow$	VICE $\uparrow$
ViT-Base	ImageNet	Supervised	–	91.238/81.657	**10.313**	0.389
ViT-Base	ImageNet	DINO	–	82.963/71.667	14.126	0.412
ViT-Base	ImageNet	MoCov3	–	69.940/53.662	15.116	**0.456**
ViT-Base	VOC2012Aug	MoCov3	–	50.173/27.923	25.317	0.331
ViT-Base	ImageNet	Supervised	*head*	88.329/80.267	**11.300**	0.430
ViT-Base	ImageNet	DINO	*head*	87.991/79.440	11.607	0.466
ViT-Base	ImageNet	MoCov3	*head*	70.966/55.786	13.107	**0.493**
ViT-Base	VOC2012Aug	MoCov3	*head*	52.738/33.101	23.105	0.404

Tables 3 and 4 reports the performance under different pre-training paradigms when using ResNet-50/ViT-Base for feature extraction. Performance and $\mathrm{ECE}_{\mathrm{ML}}$ of the models obtained by the supervised pre-training approach were superior to the self-supervised groups. However, in the experimental group loaded with the self-supervised pre-trained model, the VICE scores were generally better than those in the supervised group, which shows the potential of unsupervised pre-training in improving model interpretability.

For the evaluation protocol that fixes the backbone weights and learns only on the fc or *head* layer, the linear combination of only the features learned from pre-training stage can better exploit the advantages of different pre-training approaches. MoCov3 [4] obtains the highest VICE score in the unsupervised approach, which indicates that its learned features have strong visual interpretability. The absolute performance using the unsupervised pre-training approach is not very high, but they have a higher VICE score, indicating that their performance is more consistent with visual interpretability.

Discussion. Experimental exploration of model interpretability with various pre-training settings shows *the potential of self-supervised learning to improve the consistency between model performance and interpretability,* freeing us from the constraints of classification-based pre-training on ImageNet, allowing performance and interpretability to go hand in hand, and truly exploiting the potential of deep learning techniques. Compared to the supervised experimental group, the self-supervised schemes can achieve higher VICE score with the same settings.

4.5 How Regularization Enforces Interpretability Conformance?

Motivation. There is some recent work aimed to make classic Convnets like VGG [25] and ResNet [14] great again [6,23,30]. The performance of CNN is improved from the perspective of data and model[16,17], respectively. This indicates that the upper bound of deep neural networks has not been fully explored until now.

Table 5. The Impact of Regularization Techniques.

Model	Regularization	mAP (%) ↑/Acc (%) ↑	$\mathrm{ECE}_{\mathrm{ML}}$ (%) ↓	VICE ↑
ResNet-50	–	87.907/80.071	13.723	0.302
ResNet-50	Weight Decay	87.470/78.617	14.401	0.315
ResNet-50	Random Erasing	87.851/79.699	**13.067**	**0.323**
ViT-Base	–	90.968/82.724	9.368	0.379
ViT-Base	Weight Decay	91.238/81.657	10.313	**0.389**
ViT-Base	Random Erasing	90.765/83.190	**8.201**	0.384

Random erasing data augmentation [33] and weight decay are evaluated the performance in Table 5. Reference [19] is referred for the implementation of the weight decay regularization mechanism. We find that training with weight decay harms $\mathrm{ECE_{ML}}$, but it improves the VICE metric. Weight decay seems to improve the quality of CAMs (JI Score) by constraining the model complexity and mitigating the model overfitting. Random erasing is a data augmentation method that constructs new samples by randomly erasing a region to prevent the model from overfitting the dataset. The random erasing data augmentation avoids overconfidence and effectively improves model consistency between interpretability and performance on the VICE metrics.

Discussion. As an essential tool to avoid overfitting, the regularization not only improves the generalization performance but also impacts the interpretability of the model. It encourages the model to achieve a trade-off between performance and interpretability through extra learning objectives.

5 Other Related Work

Predictions from deep neural networks frequently suffer from over-confidence, which leads to user distrust. Accurate estimation of prediction uncertainty (model calibration) is essential for the safe application of neural networks. [24] explore regularizing neural networks by penalizing low entropy output distributions as a strong regularizer. Extensive experiments have shown in [20] that structure is a major determinant of calibration characteristics.

Many strategies have been proposed to realize model calibration based on ECE, such as temperature scaling [11], predictions ensemble [15,29]. It has been applied in model calibration for such as graph neural networks [27] and long-tailed recognition [32].

In contrast, the proposed consistency between model performance and visual interpretability VICE also deserves attention. The community should not only focus on model performance but also on achieving a better tradeoff in model performance and interpretability.

6 Conclusion

We proposed VICE as a novel sample-wise fine-grained metric to synergistically evaluate the consistency between model performance and visual interpretability from a different perspective. We selected five types of heterogeneous models for case studies using VICE and $\mathrm{ECE_{ML}}$ from different perspectives, and we found a series of inspiring phenomena:

- Blindly increasing the model capacity may harm model interpretability.
- ResNeSt and ViT, designed upon the visual attention mechanism, have better visual interpretability and $\mathrm{ECE_{ML}}$, but worse consistency between performance and visual interpretability.
- The self-supervised pre-training setting gets rid of the supervised objectives, and learns more interpretable feature representation, which shows the potential of the self-supervised learning paradigm.

- The learning objectives, introduced by different regularization, would also encourage the model to learn more interpretable features.

The proposed metrics and the found phenomena will guide us to focus on and leverage more significant impact factors for designing better performed and more interpretable models in future studies.

Acknowledgements. This work is partially supported by National Science Foundation of China (62272229), National Key Research and Development Program of China (No. 2021ZD0113200).

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
2. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: CVPR (2021)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
6. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: CVPR (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
8. Gao, S., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.H.: Res2net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2019)
9. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: PAKDD (2004)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z., Azar, M., et al.: Bootstrap your own latent: a new approach to self-supervised learning. In: NeurIPS (2020)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
12. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
16. Liang, D., Du, Y., Sun, H., Zhang, L., Liu, N., Wei, M.: Nlkd: using coarse annotations for semantic segmentation based on knowledge distillation. In: ICASSP (2021)

17. Liang, D., Geng, Q., Wei, Z., Vorontsov, D.A., Kim, E.L., Wei, M., Zhou, H.: Anchor retouching via model interaction for robust object detection in aerial images. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2021)
18. Liang, D., Li, L., Wei, M., Yang, S., Zhang, L., Yang, W., Du, Y., Zhou, H.: Semantically contrastive learning for low-light image enhancement. In: AAAI (2022)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
20. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. In: NeurIPS (2021)
21. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
22. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: AAAI (2015)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS (2019)
24. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLR (2017)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
27. Wang, X., Liu, H., Shi, C., Yang, C.: Be confident! towards trustworthy graph neural networks via confidence calibration. In: NeurIPS (2021)
28. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: CVPR (2021)
29. Wen, Y., Tran, D., Ba, J.: Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In: ICLR (2019)
30. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: an improved training procedure in timm (2021). arXiv:2110.00476
31. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: split-attention networks (2020). arXiv:2004.08955
32. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: CVPR (2021)
33. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
34. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)