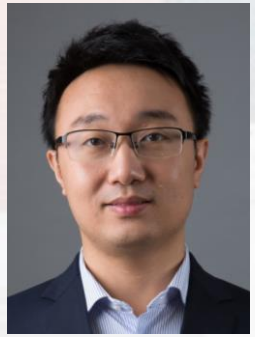
Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding







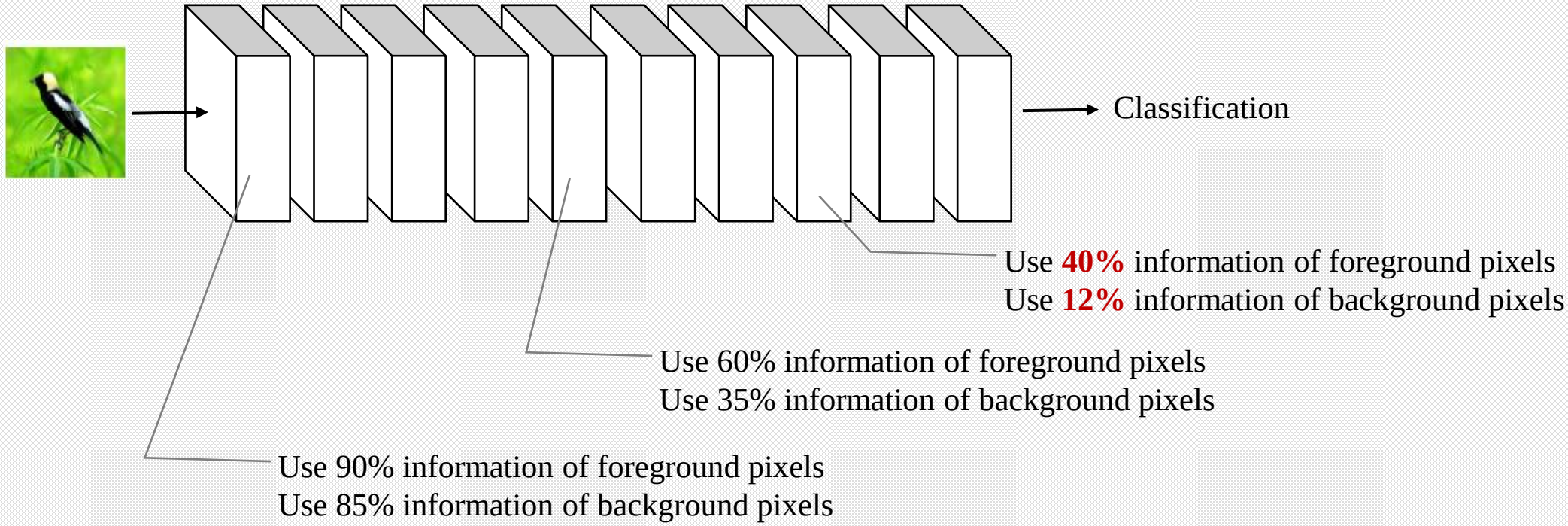
Haotian Ma[*2], Hao Zhang[*1], Fan Zhou[1], Yinqing Zhang[1], Quanshi Zhang[†1]

1. Shanghai Jiao Tong University
2. Southern University of Science and Technology
* Equal contribution † Corresponding author

Overview of the task

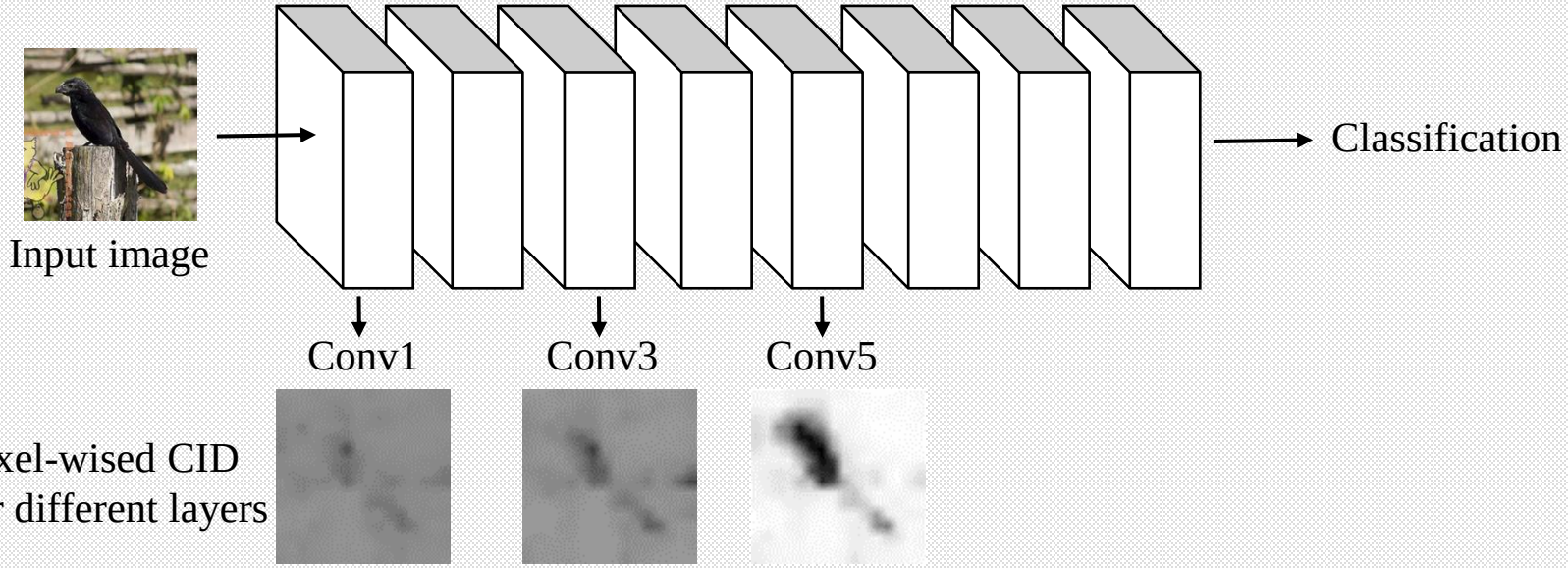
- **Motivation:**

We aim to explain how the information of each input variable is gradually discarded during the forward propagation.



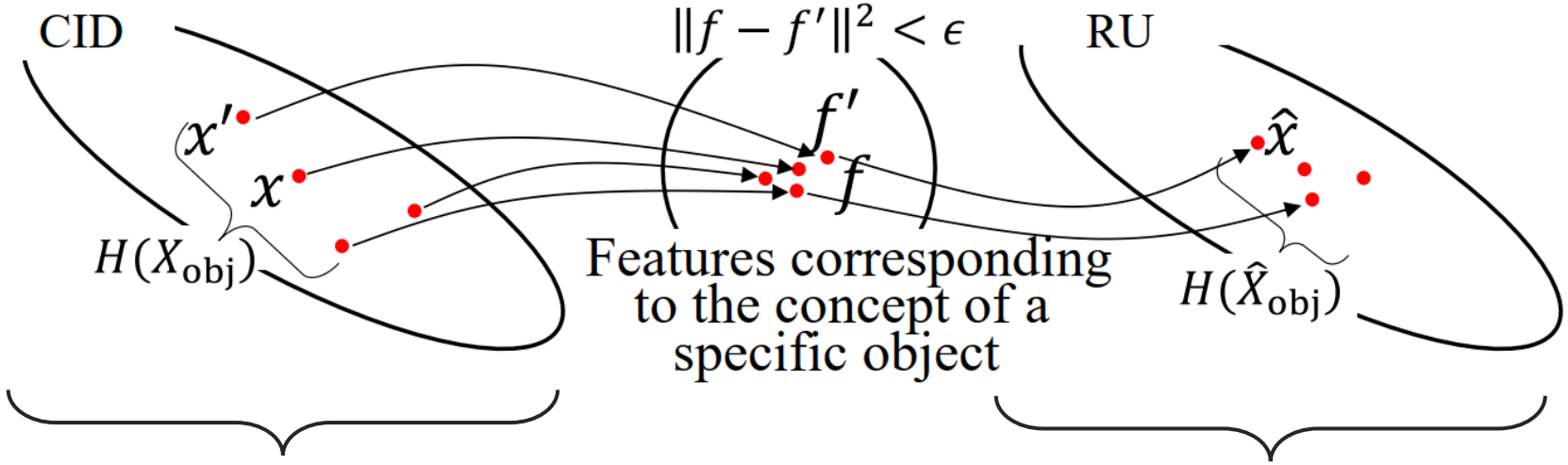
Overview of the task

- **Input:** A pre-trained neural network
  An input image
- **Output:** Quantification results of layer-wise and pixel-wise information discarding
  - How much information of each input pixel is used to compute the feature
  - How much input information can be recovered from the feature
- **Enabling fair comparability over different layers, or over different DNNs.**
- **Our metrics also have theoretical connection with the performance of DNNs.**



Metrics: CID, concentration and RU

- **Assumption:** Features in a very limited range represent a same object instance.



CID: The entropy of the perturbed input.

RU: The entropy of the reconstructed input.

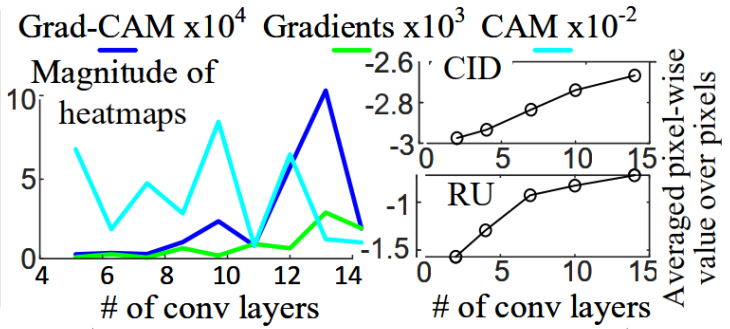
Concentration: The relative amount of information discarded in the background w.r.t. information discarded in the foreground. **A high value of the concentration indicates that the representation is effective.**

$$H(X_{obj}) = -\sum_{x' \in X_{obj}} p(x') \log p(x') \quad \text{s.t.} \quad \text{Prob}(\|f - f'\|^2 \leq \epsilon) = 1 - \tau$$

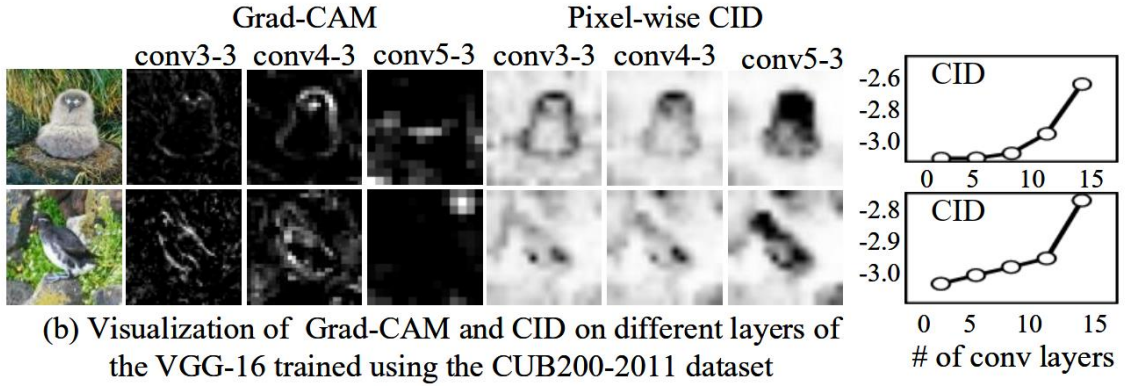
Fair comparability of the proposed metrics

	Comparability	
	Layers	Nets
Based on gradient	No	No
Based on perturbation	No	No
Based on CAM	No	No
Ours	Yes	Yes

(a1) Comparability of different explanations



(a2) Layer-wise changes of attribution maps



(b) Visualization of Grad-CAM and CID on different layers of the VGG-16 trained using the CUB200-2011 dataset

Other methods: **Cannot ensure fair comparisons** between different layers

CID and RU: **Can ensure fair comparisons** between different layers

CID and RU can be used to **fairly compare the representation capacity** between

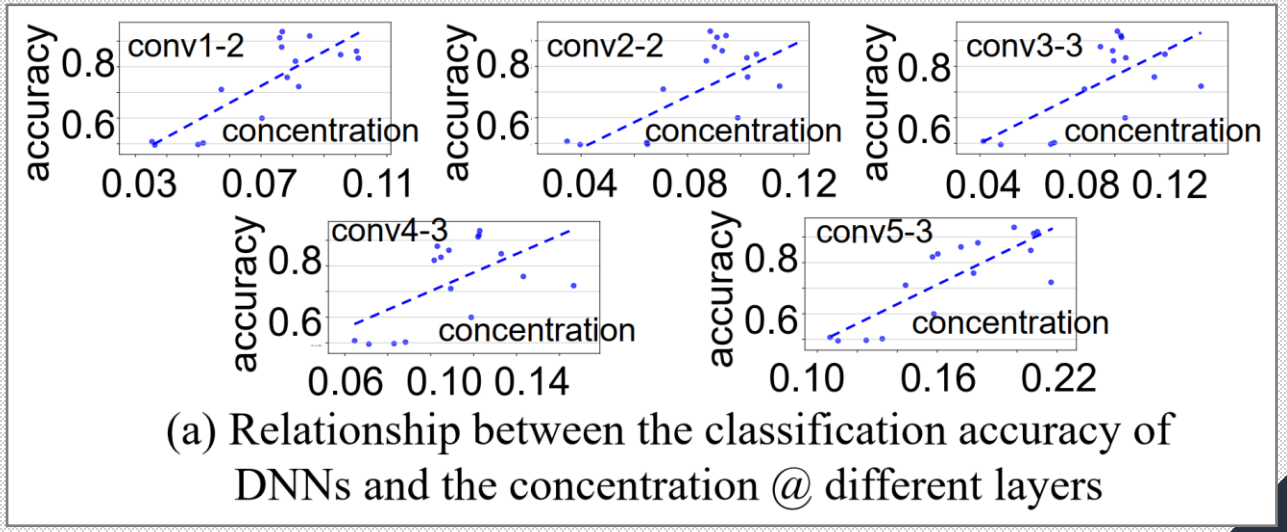
- different DNNs
- different layers of the same DNN

Positive correlation between concentration and the DNN's performance

- We theoretically proved the **positive correlation between the metric *concentration* and the efficiency of information processing** (from the perspective of **information bottleneck theory**).

$$\rho = \frac{I(F;Y)}{I(X;F)} = C_1 + \frac{C_2 concentration - C_3}{2[C_4 - CID]}$$

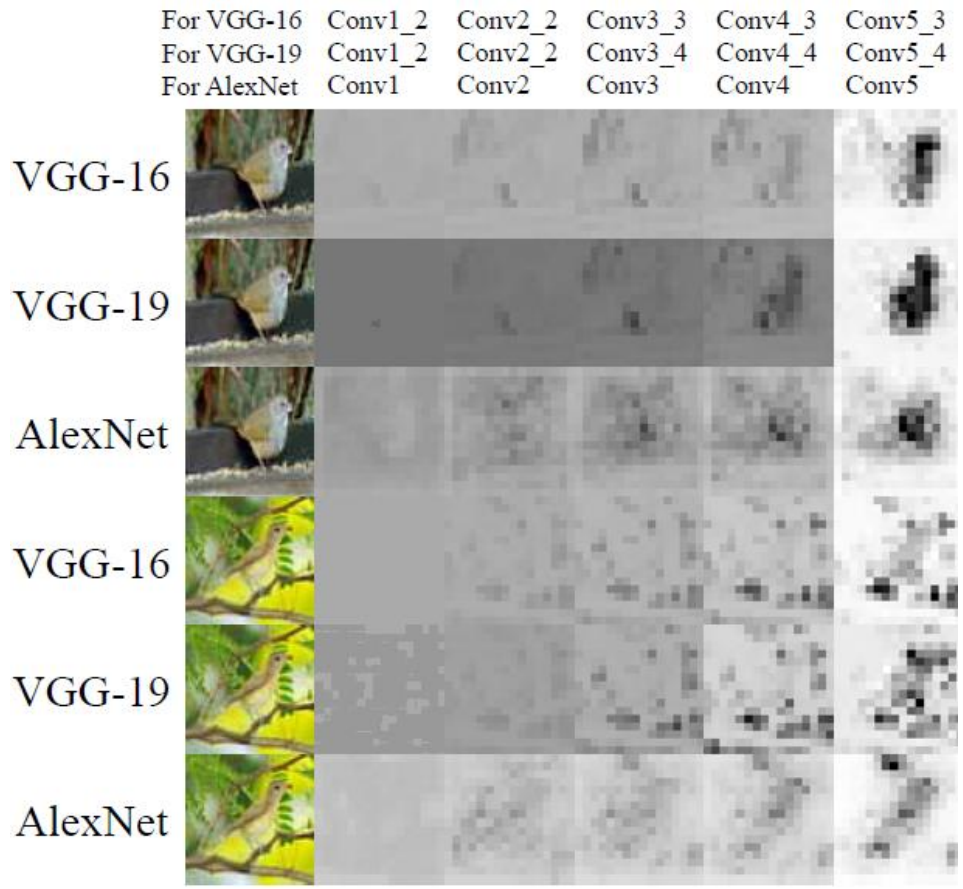
- Experimental results verified the **connection between our metric *concentration* and the accuracy of the DNN.**



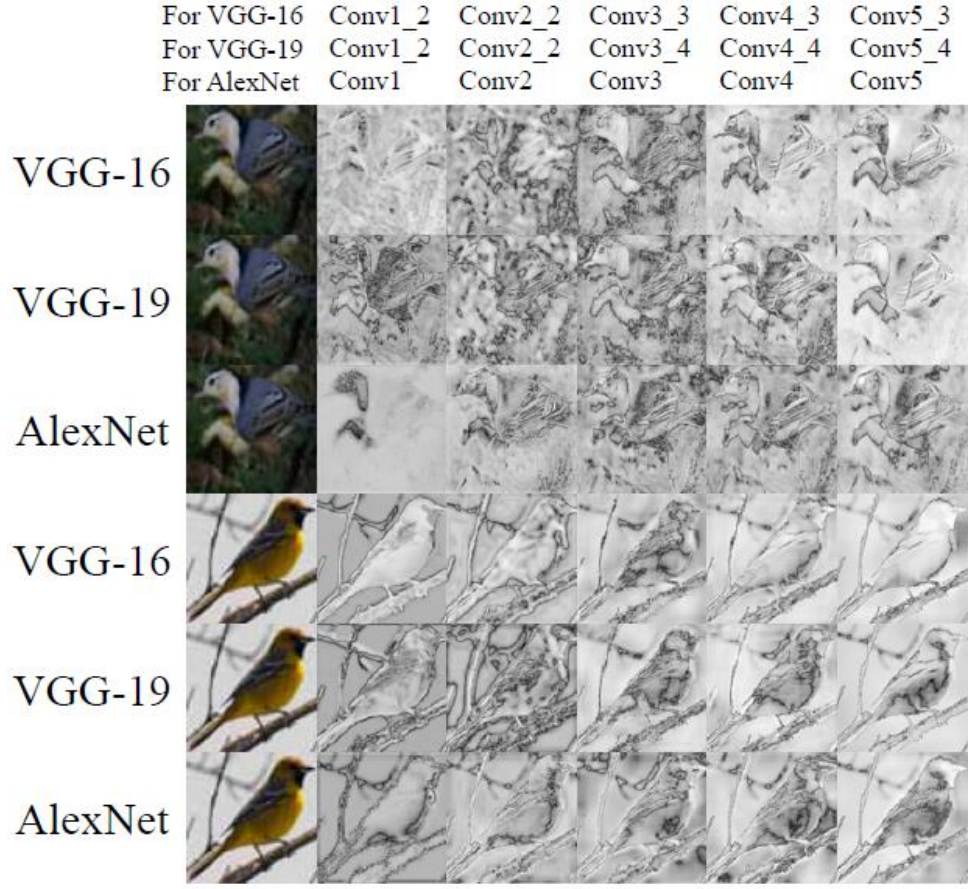
(a) Relationship between the classification accuracy of DNNs and the concentration @ different layers

Visualization of pixel-wised CID and RU

Visualization of CID



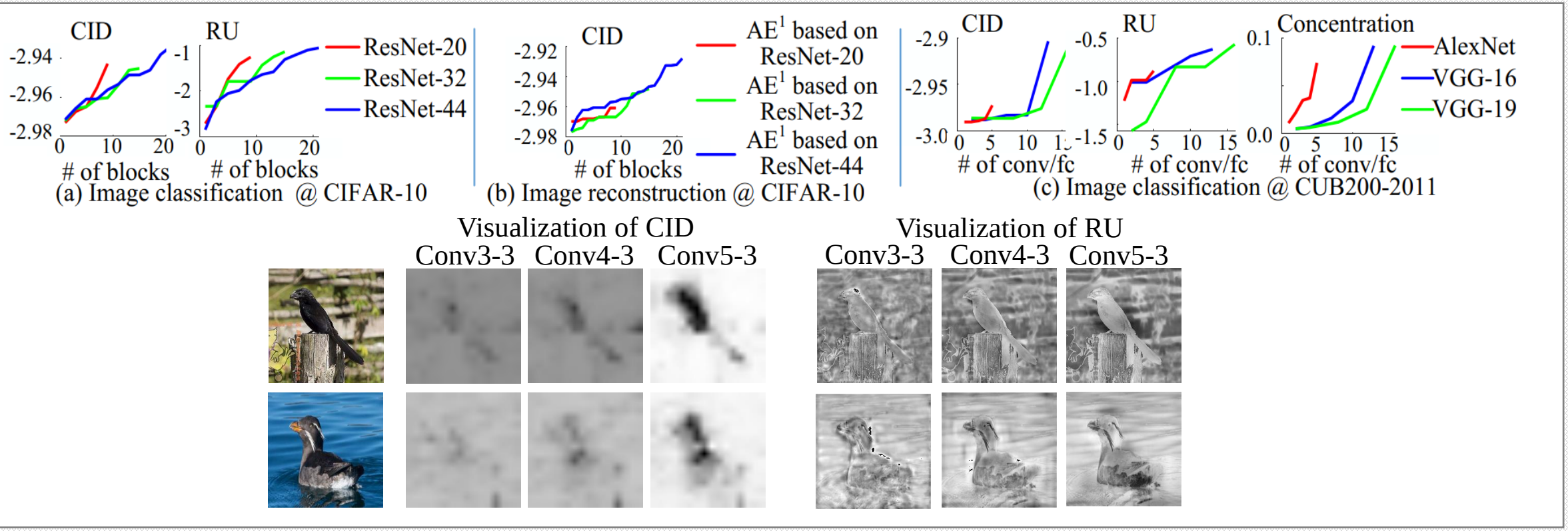
Visualization of RU



CID: DNNs discard more information in the background than information in the foreground.

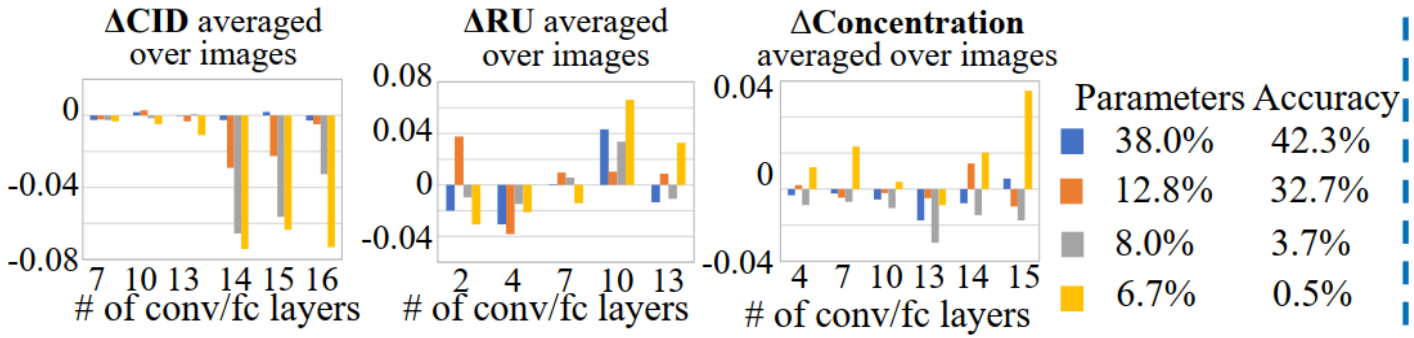
RU: Information of edges is less discarded than information of colors.

Analysis of different DNNs

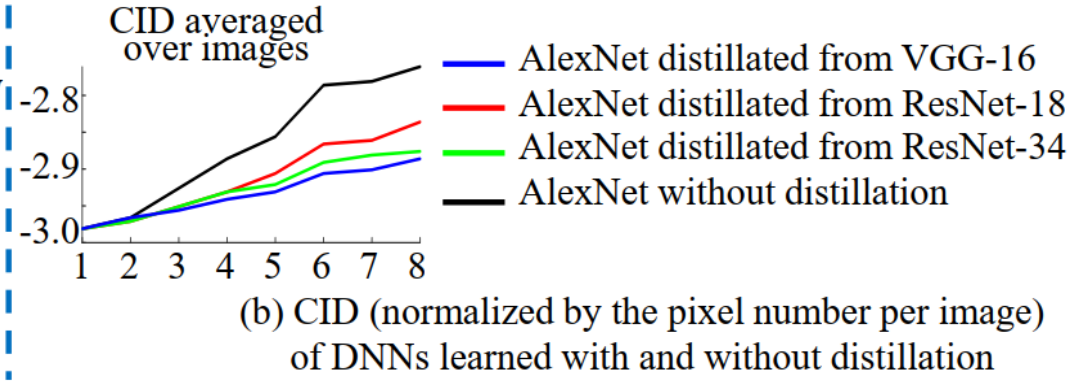


- A deep DNN usually had higher CID and RU values than a shallow DNN. Thus, **a deep DNN usually discards more input information than a shallow DNN**.
- **High layers** can be **more concentrated** on the foreground than low layers.

Analysis of existed deep learning techniques



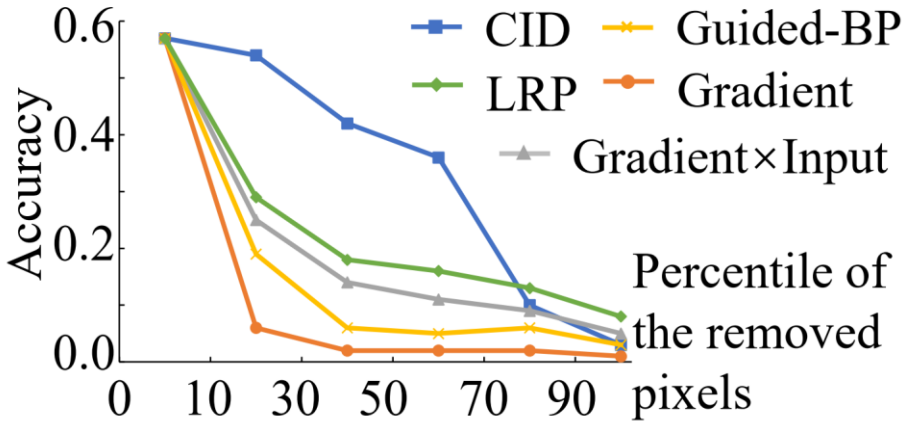
(a) Diagnosis of network compression



(b) CID (normalized by the pixel number per image) of DNNs learned with and without distillation

- **Network compression** made the DNN **less powerful to remove the information** of redundant pixels, but it still maintained the representation power of the DNN.
- **Knowledge distillation** helped the DNN to **preserve more information**.

Comparisons with previous methods



Model	ResNet-50	VGG-16
CAM	0.367	-
Grad-CAM	0.355	0.507
LRP	-	0.489
CID	**0.493**	**0.578**

- We evaluated our metric CID using the descriptive accuracy [1], and found that CID outperformed other methods.
- The metric CID outperformed CAM, Grad-CAM and LRP in the weakly-supervised localization task.

[1] Warnecke, A., Arp, D., Wressnegger, C., and Rieck, K. Evaluating explanation methods for deep learning in security. *European Symposium on Security and Privacy*, 2020.

Contributions

- We proposed metrics CID, concentration, and RU, to measure the **discarding of input information** during the forward propagation.
- Our metrics enabled **fair comparisons of the representation capacity** over different layers and different DNNs.
- In particular, we proved that **the metric concentration can reflect the efficiency of information processing.**
- Based on the proposed metrics, we analyzed **classic DNNs and existed deep learning techniques**, such as the network compression and the knowledge distillation.

THANK YOU !