

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

Haotian Ma^{*2} Hao Zhang^{*1} Fan Zhou¹ Yinqing Zhang¹ Quanshi Zhang^{†1}

Abstract

This paper presents a method to explain how the information of each input variable is gradually discarded during the forward propagation in a deep neural network (DNN), which provides new perspectives to explain DNNs. We define two types of entropy-based metrics, *i.e.* (1) the discarding of pixel-wise information used in the forward propagation, and (2) the uncertainty of the input reconstruction, to measure input information contained by a specific layer from two perspectives. Unlike previous attribution metrics, the proposed metrics ensure the fairness of comparisons between different layers of different DNNs. We can use these metrics to analyze the efficiency of information processing in DNNs, which exhibits strong connections to the performance of DNNs. We analyze information discarding in a pixel-wise manner, which is different from the information bottleneck theory measuring feature information *w.r.t.* the sample distribution. Experiments have shown the effectiveness of our metrics in analyzing classic DNNs and explaining existing deep-learning techniques. *The code is available at <https://github.com/haotianSustc/deepinfo>.*

1. Introduction

The interpretability of DNNs has received increasing attention in recent years. To this end, many methods have been proposed to measure the importance/saliency/attribution score of each input variable (Selvaraju et al., 2017; Simonyan et al., 2013; Shrikumar et al., 2016; Shapley, 1953;

^{*}Equal contribution ¹Shanghai Jiao Tong University, Shanghai, China ²Southern University of Science and Technology, Shenzhen, China. Quanshi Zhang <qs Zhang@sztu.edu.cn> is the corresponding author. He is with the Department of Computer Science and Engineering, the John Hopcroft Center and the MoE Key Lab of Artificial Intelligence, AI Institute, at the Shanghai Jiao Tong University, China.

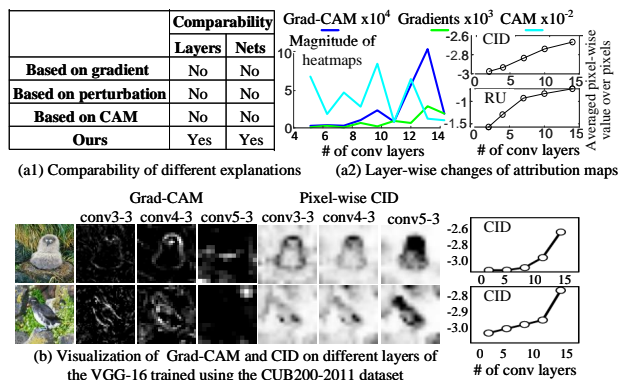


Figure 1. Subfigure (a1, a2) shows that our metrics (CID and RU) enable the fair comparison of the representation capacity of different layers. In comparison, the magnitude of explanations from previous methods is not comparable through different layers (see a2). More analysis and proof are presented in Section 3.4. Subfigure (b) visualizes the CID of each input pixel and the Grad-CAM at different layers. Appendix I has also shown importance maps generated by CAM, Gradient, and Grad-CAM on different layers of the DNN.

Springenberg et al., 2014; Lundberg & Lee, 2017; Shrikumar et al., 2016; Ribeiro et al., 2016; Fong & Vedaldi, 2017; Selvaraju et al., 2017; Zhou et al., 2016). However, these attribution maps lack the ability to reflect the representation capacity of intermediate-layer features. For example, Figure 1(a2) shows that the magnitudes of these attribution maps among different layers are quite unstable, and therefore cannot objectively reflect layer-wise changes of the representation capacity of intermediate-layer features.

Therefore, instead of merely studying the attribution score, in this paper, we aim to quantify the representation capacity of intermediate features, which provides new insight into the explanation of DNNs. To this end, the core challenge of quantifying the representation capacity is to ensure **the fair comparability of the representation capacity over different layers of the same DNN, or even over different DNNs**. We have shown that previous explanation methods cannot ensure the fairness of comparisons in Section 3.4. Therefore, in this study, we aim to explain how the information of each input variable is gradually discarded by intermediate-layer features during the forward propagation. Unlike previous studies, information discarding provides

the following new perspectives to fairly compare features through different layers in DNNs.

1. Quantification of the pixel-wise information discarding:

In general, information propagation through the cascaded layers of a DNN can be considered as a process of information selection. Figure 1(b) shows that as the number of layers increases, the DNN discards more information.

2. Efficiency of information processing: Based on our metrics, we develop a new method to quantify the efficiency of information processing of DNNs. Here, the efficiency refers to the efficiency of feature extraction of DNNs. For example, Figure 1(b) shows that from low layers to high layers, the DNN gradually shifts the attention from low-level concepts (edges) to middle-level concepts (parts), and to high-level concepts (objects).

3. Analysis of classic DNNs and classic deep-learning methods: We use our metrics to evaluate the representation capacity of classic DNNs, and analyze the effectiveness of network compression and knowledge distillation.

Metrics: To quantify the discarded information of input variables, we design two new metrics as follows.

(1) The first metric aims to quantify how much information of each input pixel is used to compute the feature, namely *pixel-wise computational information discarding* (pixel-wise CID). The information discarding refers to the phenomenon that a DNN usually selectively discards redundant information of input units (*e.g.* some pixels are not related to the task) when computing the intermediate-layer feature representation. Recently, Guan et al. (2019) proposed a method to estimate the information discarding of words in natural language processing. In this work, we extend the information discarding to the CID metric to quantify the discarded information of input pixels, and boost the fairness of layer-wise comparisons.

More crucially, based on the pixel-wise CID, we further develop a metric, namely *concentration* to measure the efficiency of the information processing of a DNN. The concentration measures the relative magnitude of information discarding on the foreground *w.r.t.* that on the background. *We theoretically explain and experimentally verify the relationship between the concentration metric and the efficiency of the information processing of the DNN (see Figure 3(a)).*

(2) The second metric aims to quantify how much input information can be recovered from the intermediate-layer feature, which is termed *pixel-wise reconstruction uncertainty* (pixel-wise RU). The RU handles the following case. Some pixels may be discarded during the forward propagation, but their information can still be well recovered by other pixels due to information redundancy.

Analysis of DNNs and findings: Unlike previous pixel-wise attribution metrics, the generality of the proposed met-

rics CID, RU, and concentration enables us to fairly compare DNNs, *i.e.* fairly compare intermediate-layer features (1) between different DNNs, and (2) between different layers of the same DNN, as Figure 1 shows. It is because our metrics are all formulated in the form of entropy, which is a generic metric in information theory, and enables fair comparisons of the DNN’s representation capacity. Furthermore, based on the metrics, we obtain the following finding.

Finding 1: The last paragraph of Section 3.1 proves a close relationship between the concentration and the DNN’s performance.

Finding 2: Network compression makes the DNN less powerful to remove the information of redundant pixels, but it still maintains the representation power of the DNN, *i.e.* the feature can still well reconstruct the input. On the other hand, the feature still concentrates on the foreground.

Finding 3: Knowledge distillation helps DNNs to preserve more information.

Besides, Appendix F also shows the proof of the relationship between the CID value and the adversarial noise.

Connection to the information bottleneck theory: The information bottleneck theory (Wolchover, 2017; Schwartz-Ziv & Tishby, 2017; Tishby & Zaslavsky, 2015) quantifies the layer-wise feature information $I(X; F)$ and $I(F; Y)$ at the *sample level*, where X represents input samples, Y represents ground-truth labels, and F denotes intermediate-layer features. In comparison, our method measures fine-grained, **pixel-wise** information discarding through layer-wise propagation. More interestingly, we prove that the metric can represent the sample-wise efficiency of feature extraction *w.r.t.* $I(X; F)$, *i.e.*, $I(F; Y)/I(X; F)$.

Contributions of this study can be summarized as follows. In this study, we propose metrics CID, concentration, and RU, to measure the discarding of input information during the forward propagation, in order to quantify the representation capacity between intermediate-layer features in a DNN. Our metrics enable fair comparisons of the representation capacity between different layers in different DNNs. Based on the proposed metrics, we analyze classic DNNs and deep learning techniques. Experiments have demonstrated the effectiveness of our method.

2. Related work

Explaining DNNs visually or semantically: The visualization of DNNs is the most direct way of explaining knowledge hidden inside a DNN (Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015; Dosovitskiy & Brox, 2016; Zhou et al., 2015; Bau et al., 2017; Fong & Vedaldi, 2018). Beyond visualization, attribution methods (Simonyan et al., 2013; Selvaraju et al., 2017; Fong & Vedaldi, 2017; Binder

Table 1. Comparisons of objectives of different explanation methods. Unlike previous methods, our method aims to quantify the representation capacity of DNNs.

Objective	Methods
Feature importance	CAM, Grad-CAM
Pixel attribution	LRP, Shapley value, SHAP, LIME, Gradient, Guided-BP
Information discarding	Our method

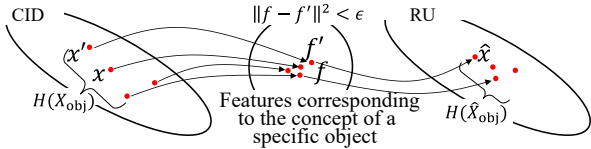


Figure 2. Illustration of the computation of CID and RU. Given a trained DNN, we compute the maximal entropy of the input $H(X_{\text{obj}})$ and the maximal entropy of image reconstruction $H(\hat{X}_{\text{obj}})$, when we constrain the intermediate-layer feature f within a small range to represent a specific object instance.

et al., 2016; Ribeiro et al., 2016; Lundberg & Lee, 2017; Springenberg et al., 2014; Zhou et al., 2016) estimated image regions that directly contribute to the network output. As Table 1 shows, our research has an essential difference from previous attribution methods. We propose to use information discarding to analyze the representation capacity of the DNN and explain classic deep learning techniques. More crucially, we prove the close relationship between our metric and the information processing of the DNN.

Mathematical evaluation of the representation capacity:

Formulating and evaluating the representation capacity of DNNs is another emerging direction. The analysis of representation similarity between DNNs based on canonical correlation analysis is widely used to analyze DNN representations (Kornblith et al., 2019; Raghu et al., 2017; Morcos et al., 2018). Novak et al. (2018) measured the sensitivity of network outputs *w.r.t.* parameters of neural networks. Zhang et al. (2017) discussed the relationship between the parameter number and the generalization capacity of DNNs. Network-attack methods (Koh & Liang, 2017) could also be used to evaluate representation robustness by computing adversarial samples for a CNN. Schulz et al. (2020) and Taghanaki et al. (2019) analyzed the feature processing from the intermediate layer to the final output of the DNN, and computing attention on intermediate layers. In comparison, this paper focuses on the processing from the input to the intermediate layer.

In particular, the information-bottleneck theory (Tishby et al., 1999) provides a generic metric to quantify the information contained in DNNs. The information-bottleneck theory can be extended to evaluate the representation capacity of DNNs (Goldfeld et al., 2019; Xu & Raginsky, 2017). Achille & Soatto (2018) further used the information-bottleneck theory to revise the dropout layer in a DNN. Our study is also inspired by the information-bottleneck theory.

Unlike analyzing the final output of a DNN in (Cheng et al., 2018), we pursue new model-agnostic and task-agnostic metrics of input information to enable comparisons over different layers of networks in a pixel-wise manner.

3. Analyze feature representations of DNNs

In order to conduct comparative studies to analyze DNNs learned by various deep-learning techniques, in this section, we introduce three generic metrics, CID, concentration, and RU. Theoretically, these metrics can be applied to various tasks, but to simplify the story, we limit our discussions to the task of object classification.

The basic idea is that we represent metrics CID, concentration, and RU as the entropy of the input information, given the feature of a specific intermediate layer. In other words, the entropy measures the uncertainty of the input when the feature represents the same object instance, *i.e.* how much input information can be discarded. Let $x \in \mathbb{R}^n$ and $f = h(x) \in \mathbb{R}^m$ denote the object instance and an intermediate-layer feature of the DNN, respectively. We assume that the DNN represents a specific object instance x using a very limited range of features with an average feature f . Similarly, there exists a latent space $X_{\text{obj}} = \{x' \mid \|h(x') - f\|^2 \leq \epsilon\}$ for x that represents the same specific object, which ensures $h(x')$ to localize in the manifold of feature f , where x' represents the perturbed input around x . ϵ is a small constant. $p(x'|X = x)$ denotes the possibility of the perturbed input x' given the input x . Let $f' = h(x') \in \mathbb{R}^m$ denote the feature in the limited range. Our method can be regarded to add perturbations to the input x to approximate the domain of f' (see Equation (2)), subject to $\|f' - f\|^2 \leq \epsilon$.

Figure 2 illustrates the basic idea of the algorithm. We compute the entropy of the input (*i.e.* the CID) when the input represents the same object instance. We also use features of the object instance to reconstruct the input $\hat{x} = g(f)$ and measure the entropy of the reconstructed input (*i.e.* the RU). In this way, two types of information discarding (CID and RU) of a specific layer can be represented using the same prototype formulation, *s.t.* $\|f' - f\|^2 \leq \epsilon$, as follows,

$$H(X_{\text{obj}}) = - \sum_{x'} p(x'|X = x) \log p(x'|X = x) \quad (1)$$

3.1. CID, concentration, and efficiency

The CID quantifies the discarding of input information during the **computation** of intermediate-layer features, which is derived from the entropy in Equation (1) from the perspective of feature extraction $f = h(x)$. The core challenge is that the explicit low-dimensional manifold of features *w.r.t.* the input x is unknown. Therefore, we approximate the manifold by adding noises to the original input x . Let

x' denote new inputs around x , *i.e.* $x' = x + \Delta x$, $x' \in \mathbb{R}^n$, which satisfy $\|f' - f\|^2 \leq \epsilon$.

Although strictly speaking, pixels in an input image are not independent, similar to (Chen et al., 2019), we assume that Δx is a Gaussian noise for simplicity, thereby $x' = x + \Delta x$ can be represented as $x' \sim \mathcal{N}(\mu = x, \Sigma)$, $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$. This also relaxes the constraint to $\text{Prob}(\|f' - f\|^2 \leq \epsilon) \geq 1 - \tau$, where $\tau \ll 1$ is a tiny positive scalar. Considering the local linearity within a small feature range of ϵ and $f = h(x)$, μ can be approximated as $\mu = x$. Although different dimensions of the input can be dependent on each other, different dimensions of the added noise can be assumed to be independent of each other. Thus, we further simplify the covariance matrix as a diagonal matrix $\Sigma = \text{diag}[\sigma_1^2, \dots, \sigma_n^2]$ to ease the computation. In this way, the CID can be decomposed to pixel-wise entropy.

$$H(X_{\text{obj}}) = \sum_{i=1}^n H_i(\sigma_i), \text{ s.t. } \text{Prob}(\|f' - f\|^2 \leq \epsilon) = 1 - \tau \quad (2)$$

where $H_i(\sigma_i) = \log \sigma_i + C$, $C = \frac{1}{2} \log(2\pi e)$. The relationship between Equations (1) and (2) is discussed in Appendix A. The overall CID value $H(X_{\text{obj}})$ can be decomposed to the pixel-wise entropy (**pixel-wise CID**) $\{H_i(\sigma_i)\}$. Figure 1(b) shows such information discarding of each input pixel. A larger value of CID indicates that the DNN discards more input information during the forward propagation.

In real applications, the overall CID can be used to compare DNNs learned on the same dataset when different DNNs share the same input size n . Our method follows the **maximum-entropy principle**, which maximizes $H(X_{\text{obj}})$ subject to constraining features within the scope of a specific object instance $\|f' - f\|^2 \leq \epsilon$. *I.e.* we enumerate all perturbation directions in x' within a small variance of f' , in order to approximate the local manifold of f' . We use the Lagrange multiplier to relax Equation (2) as follows.

$$\text{Loss}(\sigma) = \frac{1}{\delta_f^2} \mathbb{E}_{f'} [\|f' - f\|^2] - \lambda \sum_{i=1}^n H_i(\sigma) \quad (3)$$

where $\sigma = [\sigma_1, \dots, \sigma_n]^\top$ is the parameter that we aim to learn. λ is a positive scalar, and $\delta_f^2 = \lim_{\xi \rightarrow 0^+} \mathbb{E}_{x' \sim \mathcal{N}(x, \xi^2 \mathbf{I})} [\|h(x') - f\|^2] / \xi^2$ is the inherent variance of intermediate-layer features, which is used for normalization. Note that δ_f^2 is only used to normalize the intermediate-layer feature, instead of normalizing the CID value. We use $x' = x + \sigma \circ \delta$, $\delta \sim \mathcal{N}(0, I)$ to simplify the computation of the gradient *w.r.t.* σ , where \circ denotes the element-wise multiplication. Equation (3) is tractable, and we can learn σ via gradient descent.

For fair layer-wise comparisons: In order to ensure fair layer-wise comparisons, we need to control the value range of the first term in Equation (3). Features of different layers need to be perturbed at a comparable level. To this end, we

use δ_f^2 to normalize the first term in Equation (3). In this way, the stop criterion of learning σ is given as

$$\min_{\sigma} \text{Loss}(\sigma) \text{ s.t. } \mathbb{E}_{\delta \sim \mathcal{N}(0, I)} [\|f' - f\|^2] \approx \beta \delta_f^2, \beta < \alpha, \quad (4)$$

where α is a positive scalar, and $0 < \beta < \alpha$ satisfies $\mathbb{E}_{\delta \sim \mathcal{N}(0, I)} [\|f' - f\|^2] \approx \beta \delta_f^2$. The value of λ in Equation (3) is slightly adjusted (manually or automatically) to make σ satisfy $\text{Prob}(\|f' - f\|^2 \leq \alpha \delta_f^2) > 1 - \tau$. Specifically, λ is determined according to the value of β , and we will discuss the value of β in Section 4. Please see Appendix B for more details about the derivation of Equation (4).

Using the metric concentration to evaluate the efficiency of information processing: Based on the CID, we design the concentration metric to evaluate the efficiency of the feature extraction of DNNs. Given an input image x containing both the target object and some background area, let Λ denote the ground-truth segment (or the bounding box) of the target object in x . $\forall i \in \Lambda$, x_i represents pixels within Λ . Thus, the concentration is formulated as follows.

$$\text{concentration} = \frac{1}{n - |\Lambda|} \sum_{i \notin \Lambda} [H_i(\sigma_i)] - \frac{1}{|\Lambda|} \sum_{i \in \Lambda} [H_i(\sigma_i)] \quad (5)$$

Ideally, a DNN for object classification is supposed to discard background information, rather than foreground information. Note that this assumption cannot be applied to tasks depending on the background. Thus, the concentration measures the relative background information discarding *w.r.t.* foreground information discarding, which reflects the efficiency of feature extraction.

Theoretical connection of the connection between the concentration and the information bottleneck theory:

The information bottleneck theory (Wolchover, 2017; Schwartz-Ziv & Tishby, 2017) formulates the relationship between the mutual information $I(X; F)$ and $I(F; Y)$, where X denotes the input samples, and Y denotes ground-truth labels. Let $\rho = I(F; Y) / I(X; F)$ denote the efficiency of the extraction of the feature F . We can prove that a high concentration usually indicates a high value of efficiency. Specifically, we can roughly consider the foreground is related to the classification, while the background is not. Based on this, we can obtain the following relationship between the concentration value and the efficiency ρ in the form of $\rho = C_1 + \frac{C_2 \text{concentration} - C_3}{2[C_4 - \text{CID}]}$, where C_1, C_2, C_3, C_4 are four constants, and $C_2 > 0$, $C_4 > \text{CID}$. Please see Appendix C for the proof and more discussions. In addition, we find that the concentration has a close relationship with the performance of DNNs. As Figure 3 (a) shows, DNNs with better performance usually had a higher concentration.

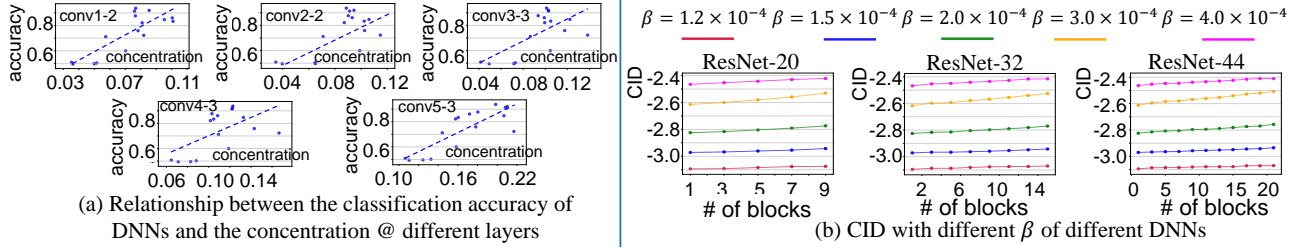


Figure 3. (a) The positive correlation between the concentration value and the classification accuracy of DNNs. Besides, the connection between the concentration and the efficiency of signal processing is explained in Appendix C. Each point corresponds to a DNN. We computed the concentration value of features in different layers of each DNN. As dashed lines show, a DNN with a high accuracy usually had a high concentration value. (b) CID computed with different values of β . A high value of β in Equation (4) led to a high value of CID. When we fixed a specific value of β , CID increased stably along with the number of blocks in the ResNet, which ensures convincing comparisons through layers.

3.2. Reconstruction uncertainty

The metric of RU is also derived from the entropy in Equation (1). The CID focuses on the input information used to compute a feature, while the RU describes the discarding of input information that can be recovered from the feature. Due to the redundancy of the input information, a pixel may be well recovered from the feature, even when the pixel is not used for feature extraction.

We use a decoder net g to reconstruct the input $\hat{x}' = g(f')$. We consider the reconstructed result \hat{x}' as the information represented by f' . Although the architecture of g affects the measurement of RU, RU values are still comparable through different layers and between DNNs when we fix g 's architecture in all comparisons. Thus, the metric RU can guarantee the fairness of layer-wise comparison (see Figure 1(a2)). Given a target DNN, g is pre-trained using the MSE loss $Loss^{dec} = \|x' - \hat{x}'\|^2$. In this way, the RU is formulated as the entropy of the reconstruction $\hat{x}' = g(f')$.

$$H(\hat{X}_{obj}) = -\sum_{\hat{x}'} p(\hat{x}') \log p(\hat{x}') \text{ s.t. } \text{Prob}(\|f' - f\|^2 \leq \epsilon) = 1 - \tau \quad (6)$$

where \hat{X}_{obj} denotes a set of images that are reconstructed using intermediate-layer features. The above entropy $H(\hat{X}_{obj})$ is computed in the same manner as the quantification of the CID. First, we synthesize the feature distribution F_{obj} by assuming that inputs follow a Gaussian distribution $x' \sim \mathcal{N}(\mu = x, \Sigma)$, $f' = h(x')$. $\hat{x}' = g(f')$ denotes the reconstructed result using f' . Second, we can also assume \hat{x}' follows a Gaussian distribution with i.i.d. random variables $\mathcal{N}(\mu^{rec} = x, \Sigma^{rec})$. As a result, the entropy of RU $H(\hat{X}_{obj})$ can be decomposed into each pixel.

$$\begin{aligned} H(\hat{X}_{obj}) &= \sum_{i=1}^n \hat{H}_i(\sigma), \quad \hat{H}_i(\sigma) = \log \hat{\sigma}_i + C \\ &= \frac{1}{2} \log \left(\mathbb{E}_{x' \sim \mathcal{N}(\mu=x, \Sigma=diag[\sigma_1, \sigma_2, \dots])} [\|\mu_i^{rec} - \hat{x}_i\|^2] \right) + C \end{aligned} \quad (7)$$

$\hat{H}_i(\sigma)$ is referred to as the **pixel-wise RU** for the i -th pixel

(unit) in the input (see Figure 4). Just like the CID, $H(\hat{X}_{obj})$ is also estimated via the maximum-entropy principle.

$$Loss(\sigma) = \frac{1}{\delta_f^2} \mathbb{E}_{f'} [\|f' - f\|^2] - \lambda \sum_{i=1}^n \hat{H}_i(\sigma)$$

We use the learned σ to compute $\hat{H}_i(\sigma)$ as the pixel-wise RU. Like the computation of CID, λ is also adjusted to ensure $\mathbb{E}_{f'} [\|f' - f\|^2] \approx \beta \delta_f^2$. The above equation is tractable and can be solved by gradient descent.

3.3. Discussions

Relationship between CID and RU: CID and RU seem to be similar metrics, but they may be significantly different in some cases. Let us consider the following two cases. (1) In the first case, redundant pixels ignored by the DNN increase the CID value, but they may still be well recovered via input reconstruction. A toy example is that given an image with a white wall, and a white pixel in the wall has the same color as its neighboring pixels. If a DNN assigns a zero weight to this pixel, then we can consider this pixel is ignored by the DNN. Thus, this pixel will have an infinite CID value. However, because this pixel and its neighboring pixels have the same color, this pixel can still be well reconstructed based on its neighboring pixels. In this case, the RU value of this pixel is still low. (2) In the second case, pixels used for feature extraction may not be reconstructed. An example is the following function $f = h(x) = \sum_i x_i$, where all pixels are used to compute the feature f . However, no pixel can be well reconstructed from $h(x)$.

Relationship between the CID and the metric in (Guan et al., 2019): Guan *et al.* (Guan et al., 2019) also measured the entropy of the input information, but there was no quantitative definition for the range of the target object. In other words, for each intermediate layer, the entropy may be measured within a different range of features, which significantly hurts the fairness of layer-wise comparisons. In comparison, we clearly define the feature range $\epsilon = \beta \delta_f^2$ to

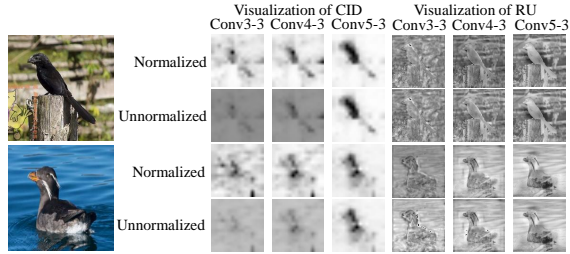


Figure 4. Visualization of pixel-wise CID and RU of different layers. The visualized pixel-wise CID and pixel-wise RU have been normalized to the value range of $[0, 1]$ to clarify the difference between foreground and background. We find that low layers mainly focus on local patterns, and high layers mainly focus on large-scale patterns. We further visualize unnormalized CID and RU values to fairly compare the information discarding between different layers. Please see Appendix G and Appendix H for more results.

enable fair layer-wise comparisons.

About information discarding in invertible networks:

Strictly speaking, there is no strict way to quantify the discarding of the input information during the computation of an intermediate-layer feature. The RU metric is related to image inversion based on invertible nets (Behrmann et al., 2019; Jacobsen et al., 2018; Kingma & Dhariwal, 2018), which also focuses on whether the feature can recover the input (theoretically, the decoder g can be implemented as the inversion operations in invertible nets). In comparison, the CID metric is defined from another perspective, *i.e.* whether the input information can contribute significant numerical values to the intermediate-layer feature or the final output. Please see Appendix D for details.

Relationship with perturbation-based methods: Our method is related to (Du et al., 2018; Fong & Vedaldi, 2017). These studies extract input pixels responsible for the intermediate-layer feature by deleting as many input pixels as possible while keeping the feature unchanged. They remove inputs by replacing inputs with human-designed values, which actually are not always meaningless. Du et al. (Du et al., 2018), Fong and Vedaldi (Fong & Vedaldi, 2017) computed pixel-wise importance. However, these methods did not enable fair comparison over layers or evaluate the representation capacity of DNNs. Please see Section 3.4 for details. In comparison, our entropy-based metrics can provide fair comparisons without specific requirements for model parameters, model architectures, and tasks.

High CID \rightarrow robustness: We can regard the forward propagation as a process of gradually discarding noisy information in the input that is irrelevant to the task, in order to extract features relevant to the task. In other words, a high CID value usually indicates that the DNN has discarded a large amount of noisy information, making the extracted features robust to noises. Specifically, people usually un-

derstand the robustness of DNNs in two aspects. The first aspect mainly considers whether the DNN’s output is largely influenced by noises, and the second aspect is whether the DNN can exhibit discrimination power on noisy samples. Strictly speaking, we can conceptually disentangle the two aspects of robustness. *I.e.*, the first aspect cares about the insensitivity to noises, and even a toy model $\forall x, h(x) = 0$ can be considered the most robust model, although it does not have any discrimination power. Whereas, the second aspect cares about the classification accuracy under noises, no matter how large the output score is changed by the noise. Appendix F.2 shows the experiment proving the relationship between the CID and the first aspect of robustness.

Limitations of concentration and RU: The concentration metric is based on the assumption that the information in the foreground is related to the task. Therefore, concentration is not suitable for tasks depending on the background. Besides, we admit that DNNs with different architectures usually need different decoder architectures. However, theoretically, our algorithm can be adapted to different decoders. Although we use the *same decoder* to fairly compare different DNNs, we still conduct experiments to test decoders with different architectures. We find that RU values do not change significantly over different decoders, which proves the trustworthiness of RU (see Appendix J for results).

Computational cost of the CID and RU is comparable with classical explanation methods, such as IG (Sundararajan et al., 2017) and LIME (Ribeiro et al., 2016). Please see Appendix E for details.

3.4. Fairness of layer-wise comparisons

In this section, we discuss the fairness of layer-wise comparisons of existing explanation metrics, as follows.

- SHAP (Lundberg & Lee, 2017) is an explanation metric based on the Shapley value (Shapley, 1953). The Shapley value directly measures the numerical contribution of each input variable to the network output, instead of the contribution to the intermediate-layer feature. Thus, the Shapley value cannot be directly used to compare the attention distribution of different layers. Besides, according to the efficiency axiom (Shapley, 1953), the sum of Shapley values of all input variables is equal to the output score. In other words, the Shapley value is sensitive to the magnitude of the network output, which disables fair comparisons between different DNNs.
- LRP (Bach et al., 2015) computes the relevance score of each variable by layer-wise relevance propagation. Compared with our metrics, the LRP mainly estimates the attention distribution over input variables, rather than explaining the information flow inside DNNs. Therefore, the LRP cannot examine the DNN’s capacity of memorizing input information.

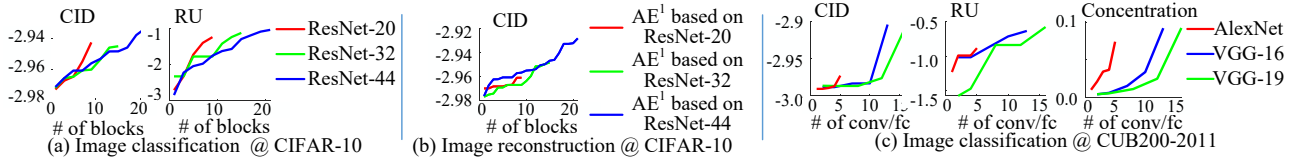


Figure 5. Layer-wise CID, RU, and concentration. Subfigure (a), (b) show that a deep DNN has high CID and RU values. Subfigure (c) shows that high layers can be more concentrated on the foreground than low layers.

• Most gradient-based methods do not generate explanations that ensures the fairness of layer-wise comparisons, such as CAM (Zhou et al., 2016), Grad-CAM (Selvaraju et al., 2017), and gradient explanations (Simonyan et al., 2013). It is because the gradient map $\frac{\partial Loss}{\partial f}$ cannot ensure the fairness of comparison between different layers. Theoretically, we can easily construct two DNNs representing exactly the same knowledge but with different magnitudes of gradients, as follows. A VGG-16 was learned to classify birds based on the CUB200-2011 dataset (Wah et al., 2011). Given a pre-trained DNN, we slightly revised the magnitude of parameters in every pair of neighboring convolutional layers $y = x \otimes w + b$ to examine our metrics. For the L -th and $(L + 1)$ -th layers, parameters were revised as $w^{(L)} \leftarrow w^{(L)}/4$, $w^{(L+1)} \leftarrow 4w^{(L+1)}$, $b^{(L)} \leftarrow b^{(L)}/4$, $b^{(L+1)} \leftarrow 4b^{(L+1)}$. Such revisions did not change knowledge representations or the network output, but changed the gradient magnitude. As Figure 1 (a2) shows, magnitudes of explanation results of baseline methods are sensitive to the magnitude of parameters. In comparison, our metrics are not affected by the magnitude of parameters, and produce reliable results. Therefore, our metrics enable layer-wise comparisons.

4. Comparative studies

We designed various experiments, in order to demonstrate the utility of the proposed metrics in comparing feature representations of various DNNs, analyzing inner mechanisms of knowledge distillation, and network compression. In order to learn the parameter σ , we used the learning rate 1×10^{-4} , and learned σ for 100 epochs.

In all experiments for image classification, we used object images cropped by object bounding boxes for both training and testing, except for experiments of computing concentration in Figure 5 where images were cropped by the box of $1.5 \text{ width} \times 1.5 \text{ height}$ of the object, which was similar to (Zhang et al., 2018). For the computation of RU, all experiments used a decoder with six residual blocks. We have tested decoders with different architectures, e.g. ResNet with different numbers of blocks. The decoder with six residual blocks had enough sophisticated architecture for feature inversion, and was relatively easy to learn. Thus, we used this decoder in experiments. To invert low-resolution features back to high-resolution images, we added two trans-

posed conv-layers to two parallel tracks in the residual block to enlarge the feature map. Considering the size of the input feature of the decoder, we added transposed conv-layers to the first 2–4 residual blocks. The effects of α and τ is controlled by β , and Figure 3 (b) shows that a low value of β led to a low value of CID. For a specific β , the CID stably increased along with the number of blocks. Therefore, the selection of β did not affect the conclusion when we used CID to analyze DNNs. In the following experiments, we set $\beta = 1.5 \times 10^{-4}$. Figure 5 visualizes the pixel-wise CID and RU for VGG-16 on the CUB200-2011 datasets. We also applied the CID to the U-Net (Ronneberger et al., 2015) trained for segmenting neuronal structures in medical images as a real-world application. The U-Net is trained using images in the ISBI cell tracking challenge (WWW, 2012), and we visualized the pixel-wised CID in Appendix G. Appendix G also shows pixel-wise CID and RU for DNNs learned on the ImageNet dataset (Russakovsky et al., 2015).

Comparisons between different DNNs for various tasks:

We compared layer-wise measures of CID and RU of different DNNs. We trained various DNNs for image classification using different datasets, and trained auto-encoders¹ (AEs) for image reconstruction (by revising architectures of ResNets-20/32/44 (He et al., 2016)). Figure 5 (a), (b) compares input information discarding of intermediate layers of both DNNs for classification and DNNs for reconstruction. We found that the CID curve of image classification and the curve of image reconstruction were similar. *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.*

Figure 5 (c) illustrates the layer-wise concentration of various DNNs, which were learned to classify birds in the CUB200-2011 (Wah et al., 2011) dataset. Compared to the AlexNet (Krizhevsky et al., 2012), we found that VGG nets (Simonyan & Zisserman, 2015) distracted attention to the background to learn diverse features in low layers, but more concentrated on the foreground object in high layers. Besides, curves became sharp at the last few layers, which indicated that *fully-connected layers made the DNN quickly discard information that is irrelevant to the task.*

¹To construct the auto-encoder, the encoder was set as all layers of the residual network before the FC layer. The decoder was the same as that for the computation of RU.

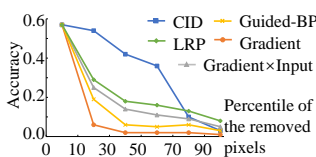


Figure 6. Accuracy of the DNN when we gradually removed pixels with the lowest importance value. A slower decrease of the accuracy indicates a higher descriptive accuracy.

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

Table 2. Accuracy of the weakly-supervised localization task. A higher value indicates a better object-localization performance.

Evaluation using the descriptive accuracy (Warnecke et al., 2020): In order to verify the effectiveness of the proposed metric, we evaluated the CID using the *descriptive accuracy* (Warnecke et al., 2020). Specifically, we gradually removed pixels with the lowest importance values, and measured the accuracy of the DNN using images after the removal. In this case, a slower decrease of the accuracy indicated a higher descriptive accuracy. We conducted this experiment with the AlexNet trained on the CUB200-2011 dataset. We compared CID with various explanation methods, including Gradient (Simonyan et al., 2013), Gradient \times Input (Shrikumar et al., 2016), Guided-BP (Springenberg et al., 2014), and LRP (Binder et al., 2016). As Figure 6 shows, the CID outperformed other methods.

Weakly-supervised localization: We further evaluated the CID via the weakly-supervised localization task (Zhou et al., 2016). We trained the VGG-16 and ResNet-50 using uncropped images from the CUB200-2011 dataset, and used ground-truth object bounding boxes for evaluation. We compared CID with several previous methods, including the CAM (Zhou et al., 2016), Grad-CAM (Selvaraju et al., 2017), LRP (Binder et al., 2016). We followed (Schulz et al., 2020) to evaluate localization results by measuring the recall rate of pixels in the bounding box, *i.e.* the number of pixels in the bounding box with high importance values. Table 4 shows the result of the evaluation. CID had a better performance than CAM, Grad-CAM, and LRP. Note that CID was not proposed to localize objects in the image. Instead, CID aimed to measure the information discarding during the forward propagation.

Analysis of network compression: We used our metrics to analyze the compressed DNN. We trained another VGG-16 using the CUB200-2011 dataset (Wah et al., 2011) for fine-grained classification. Then, the VGG-16 was compressed using the method of (Han et al., 2016) with different pruning thresholds. Figure 7 (a) compares layerwise information discarding of the original VGG-16 and the compressed VGG-16 nets with different numbers of parameters. Specifically, let $CID_{\text{compressed net}}$ and $CID_{\text{original net}}$ denote the CID value of the compressed VGG-16 and the original VGG-16, respectively. We computed the change of CID during the compression as $\Delta CID = CID_{\text{compressed net}} - CID_{\text{original net}}$. On the other hand, we also compared the re-

construction capacity and the concentration of the compressed VGG-16 and the original VGG-16. Similar to ΔCID , the change of RU and concentration is computed as $\Delta RU = RU_{\text{compressed net}} - RU_{\text{original net}}$ and $\Delta \text{concentration} = \text{concentration}_{\text{compressed net}} - \text{concentration}_{\text{original net}}$, respectively.

Based on Figure 7 (a), we found that **network compression decreased the CID of features, which indicated that compressed DNNs were more sensitive to adversarial noises**. It was because the CID value could indicate the robustness to the adversarial noise (please see Appendix F.2 for more details and results). Besides, Figure 7 shows that **network compression did not significantly affect the reconstruction capacity and the concentration of intermediate-layer features**. It meant that **the network compression made the DNN less powerful to remove the information of redundant pixels, but it still maintained the representation power of the DNN, *i.e.* the feature could still well reconstruct the input. On the other hand, the feature still concentrated on the foreground**.

Analysis of knowledge distillation: We used our metrics to analyze the inner mechanism of knowledge distillation. We trained the VGG-16, ResNet-18, and ResNet-34 using the CUB200-2011 dataset (Wah et al., 2011) as three teacher nets for fine-grained classification. Each teacher net was used to guide the learning of an AlexNet. Figure 7 (b) compares layerwise information discarding between AlexNets learned with and without knowledge distillation. We found that AlexNets learned using knowledge distillation had lower information discarding than the ordinarily learned AlexNet. Therefore, we can conclude that knowledge distillation helped AlexNets to preserve more information. Meanwhile, knowledge distillation may make intermediate-layer features more sensitive to noises, because AlexNets were mainly learned from distillation and used less noisy information from real training data during the distillation process.

Further experiments: In Appendix F, we used the proposed metrics to *analyze flaws of the network architecture, and explored the relationship between the CID value and the adversarial noises*. Furthermore, we found that *the adversarial trained DNNs discarded more information than the normally trained DNNs. Besides, the adversarial trained DNNs more focused on the foreground than the normally trained DNNs*.

5. Conclusion

In this paper, we have defined three metrics to quantify information discarding during the forward propagation. A model-agnostic method is developed to measure the proposed metrics for each specific layer of a DNN. Comparing existing methods of visualizing network features and ex-

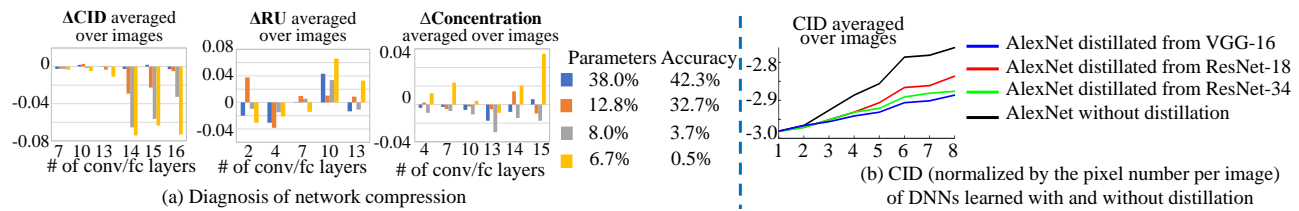


Figure 7. Analysis of network compression (a) and knowledge distillation (b). Subfigure (a) shows the change of CID, RU, and concentration after compression. We found that network compression decreased the CID, but there was no clear conclusion about the influence on the value of RU and concentration. Subfigure (b) compares layerwise information discarding between DNNs learned with and without distillations. AlexNet distilled from other DNNs discarded less information.

tracting important pixels, our metrics provide consistent and faithful results across different layers. Therefore, our metrics enable a fair analysis of the efficiency of signal processing of DNNs. The concentration value is highly correlated with the performance of the DNN. In experiments, we have used our metrics to analyze and understand the inner mechanisms of existing deep-learning techniques.

Acknowledgments This work is partially supported by National Key R&D Program of China (2021ZD0111602), the National Nature Science Foundation of China (No. 61906120, U19B2043), Shanghai Natural Science Foundation (21JC1403800, 21ZR1434600), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). This work is also partially supported by Huawei Technologies Inc.

References

- Achille, A. and Soatto, S. Information dropout: learning optimal representations through noise. *In Transactions on PAMI*, 40(12):2897–2905, 2018.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7), 2015.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. *In CVPR*, 2017.
- Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J. Invertible residual networks. *In ICML*, 2019.
- Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers. *In International Conference on Artificial Neural Networks (ICANN)*, 2016.
- Chang, B., Meng, L., Haber, E., Ruthotto, L., Begert, D., and Holtham, E. Reversible architectures for arbitrarily deep residual neural networks. *In AAAI*, 2018.
- Chen, R., Chen, H., Huang, G., Ren, J., and Zhang, Q. Explaining neural networks semantically and quantitatively. *In ICCV*, 2019.
- Cheng, H., Lian, D., Gao, S., and Geng, Y. Evaluating capability of deep neural networks for image classification via information plane. *In ECCV*, 2018.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *In CVPR*, 2009.
- Dinh, L., Krueger, D., and Bengio, Y. NICE: non-linear independent components estimation. *In ICLR*, 2015.
- Dosovitskiy, A. and Brox, T. Inverting visual representations with convolutional networks. *In CVPR*, 2016.
- Du, M., Liu, N., Song, Q., and Hu, X. Towards explanation of dnn-based prediction with guided feature inversion. *In arXiv:1804.00506*, 2018.
- Fong, R. and Vedaldi, A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *In CVPR*, 2018.
- Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *In ICCV*, 2017.
- Goldfeld, Z., van den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating information flow in deep neural networks. *In ICML*, 2019.
- Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. Analyzing inverse problems with invertible neural networks. *In ICLR*, 2019.
- Guan, C., Wang, X., Zhang, Q., Chen, R., He, D., and Xie, X. Towards a deep and unified understanding of deep neural models in nlp. *In ICML*, 2019.
- Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *In ICLR*, 2016.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *In CVPR*, 2016.

- Jacobsen, J., Smeulders, A. W. M., and Oyallon, E. i-revnet: Deep invertible networks. *In ICLR*, 2018.
- Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *In NIPS*, 2018.
- Koh, P. and Liang, P. Understanding black-box predictions via influence functions. *In ICML*, 2017.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. *In arXiv:1905.00414*, 2019.
- Krizhevsky, A. Learning multiple layers of features from tiny images. *In Technical report*, 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *In NIPS*, 2012.
- Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *In NIPS*, 2017.
- Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them. *In CVPR*, 2015.
- Morcos, A. S., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. *In NIPS*, 2018.
- Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: An empirical study. *In ICLR*, 2018.
- Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *In NIPS*, 2017.
- Ribeiro, M. T., Singh, S., and Guestrin, C. “why should i trust you?” explaining the predictions of any classifier. *In KDD*, 2016.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Schulz, K., Sixt, L., Tombari, F., and Landgraf, T. Restricting the flow: Information bottlenecks for attribution. *In International Conference on Learning Representations*, 2020.
- Schwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *In arXiv:1703.00810*, 2017.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In ICCV*, 2017.
- Shapley, L. S. A value for n-person games. *In Contributions to the Theory of Games*, 2(28):307–317, 1953.
- Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *In arXiv:1605.01713*, 2016.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *In ICLR*, 2015.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *In arXiv:1312.6034*, 2013.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *In arXiv:1412.6806*, 2014.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *In International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.
- Taghanaki, S. A., Havaei, M., Berthier, T., Dutil, F., Di Jorio, L., Hamarneh, G., and Bengio, Y. Infomask: Masked variational latent representation to localize chest disease. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 739–747, Cham, 2019.
- Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. *In arXiv:1503.02406*, 2015.
- Tishby, N., Pereira, F., and Bialek, W. The information bottleneck method. *In arXiv:physics/0004057*, 1999.
- Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical report, In California Institute of Technology, 2011.
- Warnecke, A., Arp, D., Wressnegger, C., and Rieck, K. Evaluating explanation methods for deep learning in security. *European Symposium on Security and Privacy*, 2020.

Wolchover, N. New theory cracks open the black box of deep learning. *In Quanta Magazine*, 2017.

WWW. Web page of the em segmentation challenge. http://brainiac2.mit.edu/isbi_challenge/, 2012.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *In NIPS*, 2017.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. *In ECCV*, 2014.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Undersantding deep learning requires rethinking generalization. *In ICLR*, 2017.

Zhang, Q., Wu, Y. N., and Zhu, S.-C. Interpretable convolutional neural networks. *In CVPR*, 2018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. *In ICRL*, 2015.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. *In CVPR*, 2016.

A. Relationship between Equation (1) and Equation (2)

We introduce the computation of the entropy of the input in Equation (1) of the paper, and the computation of the CID value in Equation (2) of the paper. In this section, we introduce how to approximate Equation (1) using Equation (2). For the convenience of readers, we rewrite these equations as follows.

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

$$H(X_{\text{obj}}) = \sum_{i=1}^n H_i(\sigma_i), \quad \text{s.t.} \quad \text{Prob}(\|f' - f\|^2 \leq \epsilon) = 1 - \tau$$

$x' \in X_{\text{obj}}$ is a perturbed input around the original input x , *i.e.* $x' = x + \Delta x$. We assume that x' follows the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. However, this assumption cannot ensure $\|f' - f\|^2 \leq \epsilon$, because there is a small probability that x' is significantly different from x . Therefore, we relax the constraint as $\text{Prob}(\|f' - f\|^2 \geq 1 - \tau)$, where τ is a small positive number. We simplify the covariance matrix as $\Sigma = \text{diag}[\sigma_1^2, \dots, \sigma_n^2]$, and we can approximate $H(X_{\text{obj}})$ as $\sum_{i=1}^n H_i(\sigma_i)$. In this way, we can approximate Equation (1) using Equation (2).

B. The derivation of Equation (4) in the paper

In the paper, we introduce the computation of the metric CID in the ‘‘For fair layer-wise comparisons’’ paragraph of Section 3.1. We learn the σ to compute the CID value using Equation (3). In this section, we introduce how to get Equation (4) to ensure the fairness of the layer-wise comparison. For the convenience of readers, we rewrite Equation (3) as follows.

$$\text{Loss}(\sigma) = \frac{1}{\delta_f^2} \mathbb{E}_{f'} [\|f' - f\|^2] - \lambda \sum_{i=1}^n H_i(\sigma),$$

where $\sigma = [\sigma_1, \dots, \sigma_n]^\top$ is the parameter that we aim to learn. λ is a positive scalar, and $\delta_f^2 = \lim_{\xi \rightarrow 0^+} \mathbb{E}_{x' \sim \mathcal{N}(x, \xi^2 \mathbf{I})} [\|h(x') - f\|^2] / \xi^2$ is the inherent variance of intermediate-layer features, which is used for normalization.

In order to ensure fair layer-wise comparisons, we need to control the value range of the first term in Equation (3). Features of different layers need to be perturbed at a comparable level. Therefore, for each layer, we measure and compare $H_i(\sigma_i)$ when $\|f' - f\|^2 \leq \epsilon = \alpha \delta_f^2$, where α is a positive scalar. In this way, the value of λ need to be adjusted (manually or automatically) to make σ satisfy $\text{Prob}(\|f' - f\|^2 \leq \alpha \delta_f^2) > 1 - \tau$.

To simplify the implementation, we make the approximation $\mathbb{E}_{\delta \sim \mathcal{N}(0, I)} [\|f' - f\|^2] \approx \beta \sigma_f^2$, where $\beta < \alpha$, as a replacement of $\text{Prob}(\|f' - f\|^2 \leq \alpha \delta_f^2) > 1 - \tau$. In this way, λ is determined based on the value of β , and we do not need to specify values of τ and α . *I.e.* we only need to consider the value of β to learn σ . Thus, the stop criterion of learning σ can be given as Equation (4) in the paper, *i.e.*

$$\min_{\sigma} \text{Loss}(\sigma) \quad \text{s.t.} \quad \mathbb{E}_{\delta \sim \mathcal{N}(0, I)} [\|f' - f\|^2] \approx \beta \sigma_f^2, \beta < \alpha.$$

C. Proof of the relationship between the concentration and the information bottleneck theory

We introduce the relationship between the concentration and the information bottleneck theory in the last paragraph of Section 3.1. In this section, we prove the above relationship, *i.e.* a high concentration usually indicates a high efficiency ρ . According to the information bottleneck theory (Wolchover, 2017; Schwartz-Ziv & Tishby, 2017), the efficiency ρ of a DNN can be computed as $\rho = I(F; Y) / I(X; F)$, where X, F, Y denote input samples, intermediate-layer features and ground-truth labels, respectively. The efficiency can be formulated as follows.

$$\begin{aligned} \rho &= \frac{I(F; Y)}{I(X; F)} \\ &= \frac{I(X; F; Y) + I(F; Y|X)}{H(X) - H(X|F)} \\ &= \frac{I(X; Y) - I(X; Y|F) + I(F; Y|X)}{H(X) - H(X|F)} \\ &= \frac{H(X) - H(X|Y) - I(X; Y|F) + I(F; Y|X)}{H(X) - H(X|F)} \\ &= \frac{H(X) - H(X|Y) - I(X; Y|F) + I(F; Y|X)}{H(X) - \text{CID}} \end{aligned} \tag{8}$$

Note that the intermediate-layer feature F is determined by X , thereby $I(F; Y|X) = 0$. In this way, the above equation can be rewritten as

$$\rho = \frac{H(X) - H(X|Y) - I(X; Y|F)}{H(X) - \text{CID}} \quad (9)$$

Since $H(X)$ and $H(X|Y)$ are only related to the dataset, we can consider them as constants. Given F , we assume that the foreground of the input is conditionally independent with the background. Thus, $I(X; Y|F)$ can be disentangled as $I(X; Y|F) = I(X_{\text{fg}}; Y|F) + I(X_{\text{bg}}; Y|F)$, where X_{fg} and X_{bg} denote the foreground and background part of the input, respectively. Specifically, we have

$$\begin{aligned} I(X_{\text{fg}}; Y|F) &= H(X_{\text{fg}}|F) - H(X_{\text{fg}}|F, Y) = \gamma_{\text{fg}}H(X_{\text{fg}}|F) \\ I(X_{\text{bg}}; Y|F) &= H(X_{\text{bg}}|F) - H(X_{\text{bg}}|F, Y) = \gamma_{\text{bg}}H(X_{\text{bg}}|F) \end{aligned} \quad (10)$$

We assume that there exists a scalar γ_{fg} to represent the ratio of the foreground information, which is related to the ground-truth label Y , *i.e.* $I(X_{\text{fg}}; Y|F) = \gamma_{\text{fg}}H(X_{\text{fg}}|F)$. Similarly, we assume that there exists a scalar γ_{bg} to represent the ratio of the background information, which is related to the ground-truth label Y , *i.e.* $I(X_{\text{bg}}; Y|F) = \gamma_{\text{bg}}H(X_{\text{bg}}|F)$. Since $H(X_{\text{fg}}|F) > H(X_{\text{fg}}|F, Y)$ and $H(X_{\text{bg}}|F) > H(X_{\text{bg}}|F, Y)$, we have $0 < \gamma_{\text{fg}} < 1$, $0 < \gamma_{\text{bg}} < 1$. Since the task is mainly related to the foreground, the information discarded in the foreground is usually less than the information discarded in the background. In this way, we have $\gamma_{\text{fg}} \gg \gamma_{\text{bg}}$, $\gamma_{\text{fg}} - \gamma_{\text{bg}} > 0$. Thus, the efficiency ρ can be written as follows.

$$\rho = \frac{H(X) - H(X|Y) - \gamma_{\text{fg}}H(X_{\text{fg}}|F) - \gamma_{\text{bg}}H(X_{\text{bg}}|F)}{H(X) - \text{CID}} \quad (11)$$

$$= \frac{\gamma_{\text{fg}}\text{concentration} - (\gamma_{\text{fg}} + \gamma_{\text{bg}})H(X_{\text{bg}}|F) + H(X) - H(X|Y)}{H(X) - \text{CID}} \quad (12)$$

$$= -\frac{\gamma_{\text{bg}}\text{concentration} + (\gamma_{\text{fg}} + \gamma_{\text{bg}})H(X_{\text{fg}}|F) - H(X) + H(X|Y)}{H(X) - \text{CID}} \quad (13)$$

By combining above two equations, we have

$$\begin{aligned} \rho &= \frac{(\gamma_{\text{fg}} - \gamma_{\text{bg}})\text{concentration} - (\gamma_{\text{fg}} + \gamma_{\text{bg}})\text{CID} + 2(H(X) - H(X|Y))}{2[H(X) - \text{CID}]} \\ &= \frac{\gamma_{\text{fg}} + \gamma_{\text{bg}}}{2} + \frac{(\gamma_{\text{fg}} - \gamma_{\text{bg}})\text{concentration} - (\gamma_{\text{fg}} + \gamma_{\text{bg}} - 2)H(X) - 2H(X|Y)}{2[H(X) - \text{CID}]} \end{aligned} \quad (14)$$

Note that γ_{fg} , γ_{bg} , $H(X)$ and $H(X|Y)$ can be considered as constants. For simplicity, let $C_1 = \frac{\gamma_{\text{fg}} + \gamma_{\text{bg}}}{2}$, $C_2 = \gamma_{\text{fg}} - \gamma_{\text{bg}} > 0$, $C_3 = (\gamma_{\text{fg}} + \gamma_{\text{bg}} - 2)H(X) + 2H(X|Y)$, $C_4 = H(X) > \text{CID}$. Therefore, we have

$$\rho = C_1 + \frac{C_2\text{concentration} - C_3}{2[C_4 - \text{CID}]} \quad (15)$$

Thus, for DNNs learned for the same task with the similar value of CID, a high value of concentration usually indicates a high value of efficiency ρ , which reflects the connection between our metrics and the information bottleneck theory.

D. About how to understand the limitation of CID from the perspective of invertible nets

In this section, we discuss the limitation of CID in invertible nets, which is briefly introduced in the third paragraph of Section 3.3.

Strictly speaking, there is no strict way to quantify the discarding of the input information during the computation of an intermediate-layer feature. Our method is based on the assumption that the concept of a specific object instance is within the range of $\text{Prob}(\|f' - f\|^2 \leq \epsilon) \geq 1 - \tau$, which makes the algorithm sensitive to the activation magnitude of each feature dimension. For example, a typical failure case for this assumption is invertible neural networks (Behrmann et al., 2019; Chang et al., 2018; Gomez et al., 2019; Jacobsen et al., 2018; Kingma & Dhariwal, 2018; Dinh et al., 2015). Theoretically, invertible neural networks do not discard any input information; otherwise, the input cannot be inverted from intermediate-layer features. Instead, invertible neural networks usually significantly decrease the magnitude of neural activations caused by unimportant pixels *w.r.t.* the task, and boost the magnitude of neural activations triggered by important pixels *w.r.t.* the

task. Similarly, given a pre-trained DNN, if we revise a DNN by selectively halving magnitudes of parameters of 50% filters $w \leftarrow 0.5w$, theoretically, this revision does not discard any input information.

However, information discarding in this paper is defined from another perspective, *i.e.* whether the input information can significantly contribute to the final output of the neural network. For both invertible neural networks and the above revision of halving magnitudes of parameters, these techniques all decrease activation magnitudes caused by certain pixels, thereby letting these pixels contribute less numerical values to the network output.

Therefore, our definition of information discarding does not conflict with the information processing in invertible neural networks. Based on our definition of information discarding, a high information discarding of a pixel indicates that this pixel will contribute a low numerical score to the intermediate-layer feature or the network output.

E. About the computational cost of CID and RU

In the last paragraph of Section 3.3, we have briefly clarified that the computational cost of the CID and RU is comparable with previous explanation methods. In this section, we will provide more discussions about this issue. In this paper, the pixel-wise CID and RU were usually generated by letting the DNN recursively conduct 100 inferences. In comparison, IG (Sundararajan et al., 2017) took 300 inferences to compute the attribution map. LIME (Ribeiro et al., 2016) needed 5000 inferences to learn the explanation result. The computational cost of the Shapley value (Shapley, 1953) was NP-hard. All of these explanation methods had a higher computational cost than the proposed metrics. Therefore, the

F. Further experiments

This section introduces several additional experiments, which are briefly introduced in the last paragraph of the "Comparative studies" section in the paper.

F.1. Diagnosis of architectural revision (damage)

In this experiment, we aimed to analyze whether the proposed metrics reflected architectural revisions of DNNs. To this end, we revised the architecture of the VGG-16/VGG-19 network by changing a specific convolutional layer to contain four $7 \times 7 \times 512$ filters with padding=3, which hurts the representation capacity of the DNN. We trained both the original VGG-16/VGG-19 and the revised VGG-16/VGG-19 for binary classification between bird images cropped from the CUB200-2011 dataset (Wah et al., 2011) and random images in the ImageNet (Deng et al., 2009). Figure 8 compares the original and revised DNNs. We found that compared to the original DNN, the architectural revision significantly boosted the information discarding at the revised layer. Meanwhile, the architectural revision (damage) also slightly increased the concentration of DNNs. The increase of the concentration seemed to conflict with the architectural damage, but this can be explained as follows.

1. Compared to the increase of the information discarding of the revised net, the increase of concentration was significantly lower. Thus, in general, the architectural revision hurt the representation capacity of the DNN.
2. The DNN with the reduced feature dimension could only encode much fewer concepts of object parts. Thus, the revised DNN usually encoded fewer, simpler, but more discriminative features than original DNNs.
3. Original DNNs usually ignored background information and extracted discriminative foreground features at high FC layers (see Figure 8, whereas the dimension reduction at the revised layer made the DNN ignored background information at much lower layers.

Table 3. Relationship between the ΔCID and the adversarial robustness. A higher CID value usually indicates a higher adversarial robustness.

DNN	ΔCID	adversarial robustness $\ \epsilon\ _2$
Original DNN (with 100% parameters)	0	0.00276
DNN with 38.0% parameters	-0.003	0.00281
DNN with 12.8% parameters	-0.005	0.00254
DNN with 8.0% parameters	-0.033	0.00109

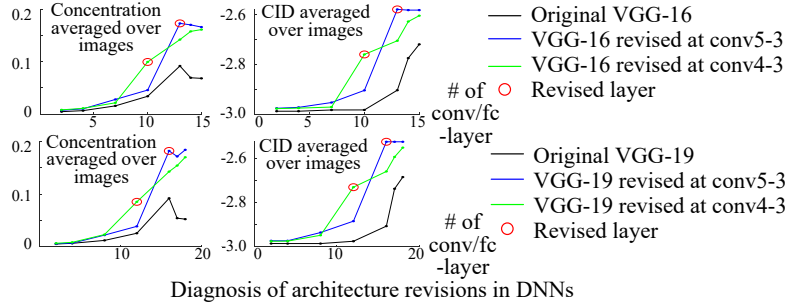


Figure 8. Diagnosis of architectural revision. Values were normalized by the pixel number per image and averaged over images. The architectural revision increased the value of CID and concentration.

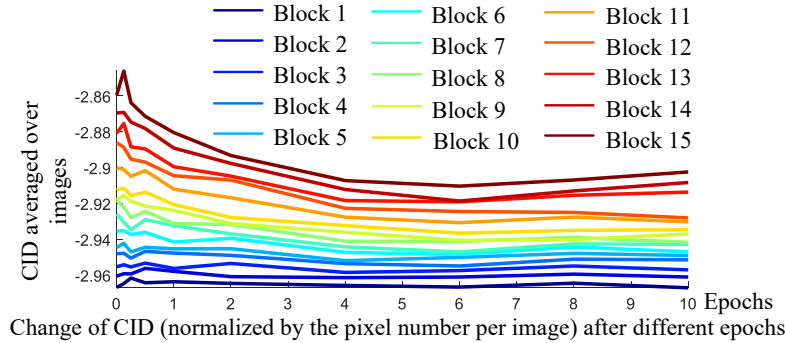


Figure 9. Effects of learning epochs. CID values were normalized by the pixel number per image and averaged over images. Each curve shows the information discarding of the output feature of a specific block during the learning process.

F.2. Relationship to adversarial noises

We conducted an experiment to reveal the relationship between the CID value and the adversarial noise of the DNN. We trained a VGG-16 using the CUB200-2011 dataset (Wah et al., 2011) for fine-grained classification. Then, the VGG-16 was compressed using the method of (Han et al., 2016) with different pruning thresholds. Table 3 compares the CID value of the last FC layer and the adversarial robustness of the DNN, when the DNN was compressed at different ratios. For each input image, we computed adversarial samples towards top-20 incorrect fine-grained categories with the highest probabilities. For fair comparisons, we added the adversarial noise until the adversarial attack just succeeded, *i.e.* when the adversarial perturbation just pushed the sample to the decision boundary. For each adversarial noise, we measured its L-2 norm values. The adversarial robustness was reported as the average L-2 norm over all images. We only measured the CID value of the last FC layer, because the CID value of the last layer most fit the logic of the final prediction. Table 3 shows that a higher CID value usually indicates a higher adversarial robustness. This indicated a close relationship between the CID value and the adversarial noise.

F.3. Analysis of information discarding after different epochs during the learning process

We trained the ResNet-32 network using the CIFAR-10 dataset (Krizhevsky, 2009). Figure 9 shows the change of information discarding *w.r.t.* output features of different blocks during the learning process. Information discarding in high layers satisfied the information-bottleneck theory.

F.4. Analysis of information discarding in the adversarial attack

In this experiment, we compared an adversarially trained AlexNet and a normally trained AlexNet on the CUB-200 2011 dataset. During the adversarial training, we used the PGD to generate adversarial samples. Specifically, we used the L-∞ attack, where the number of steps of the attack is 10, and the step size of the attack is 0.001. Figure 10 shows that the adversarial trained AlexNet discarded more information than the normally trained AlexNet, which was consistent with results in Appendix F.2. Besides, the adversarial trained AlexNet more focused on the foreground than the normally trained

AlexNet, since the adversarially trained AlexNet had higher *concentration* values.

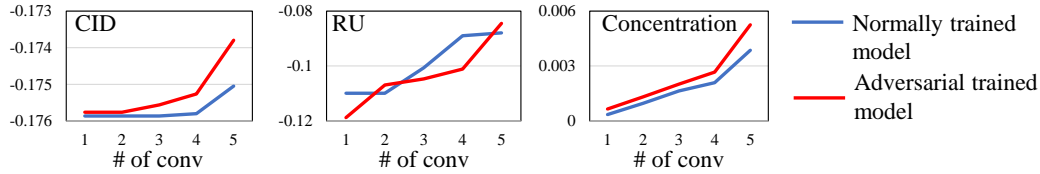
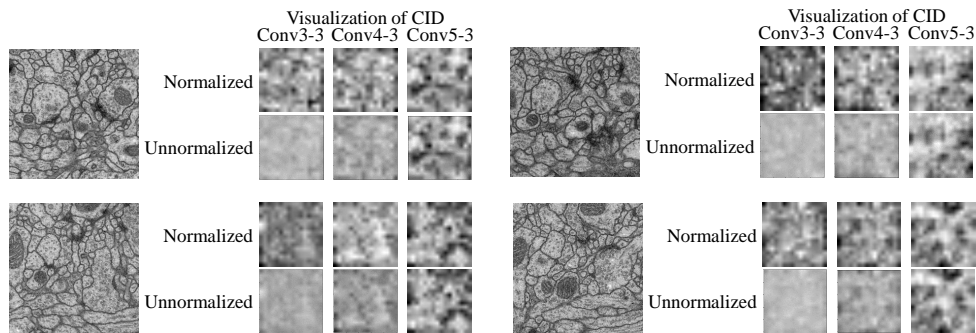


Figure 10. Comparison of the CID, RU, and *concentration* value between the adversarially trained AlexNet and the normally trained AlexNet. The adversarial trained AlexNet discarded more information than the normally trained AlexNet, and more focused on the foreground than the normally trained AlexNet.

G. Visualization of pixel-wise CID

G.1. For the U-Net learning using images in the ISBI cell tracking challenge

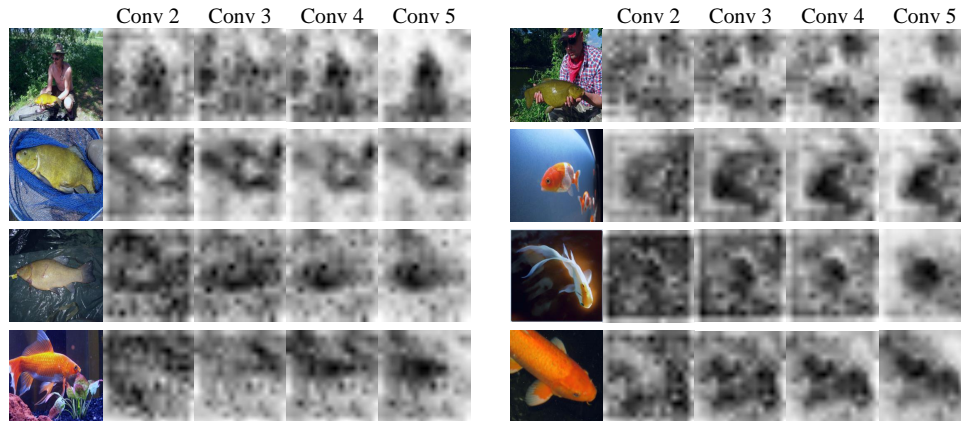
In the second paragraph of Section 4, we have claimed that we applied the metric CID to the U-Net trained for segmenting neuronal structures in medical images as a real-world application. In this subsection, we will provide visualization results of the pixel-wise CID computed on different layers of the U-Net.



Visualization of CID for U-Net learned on images in the ICBI cell tracking challenge

G.2. For the AlexNet learned using the ImageNet dataset

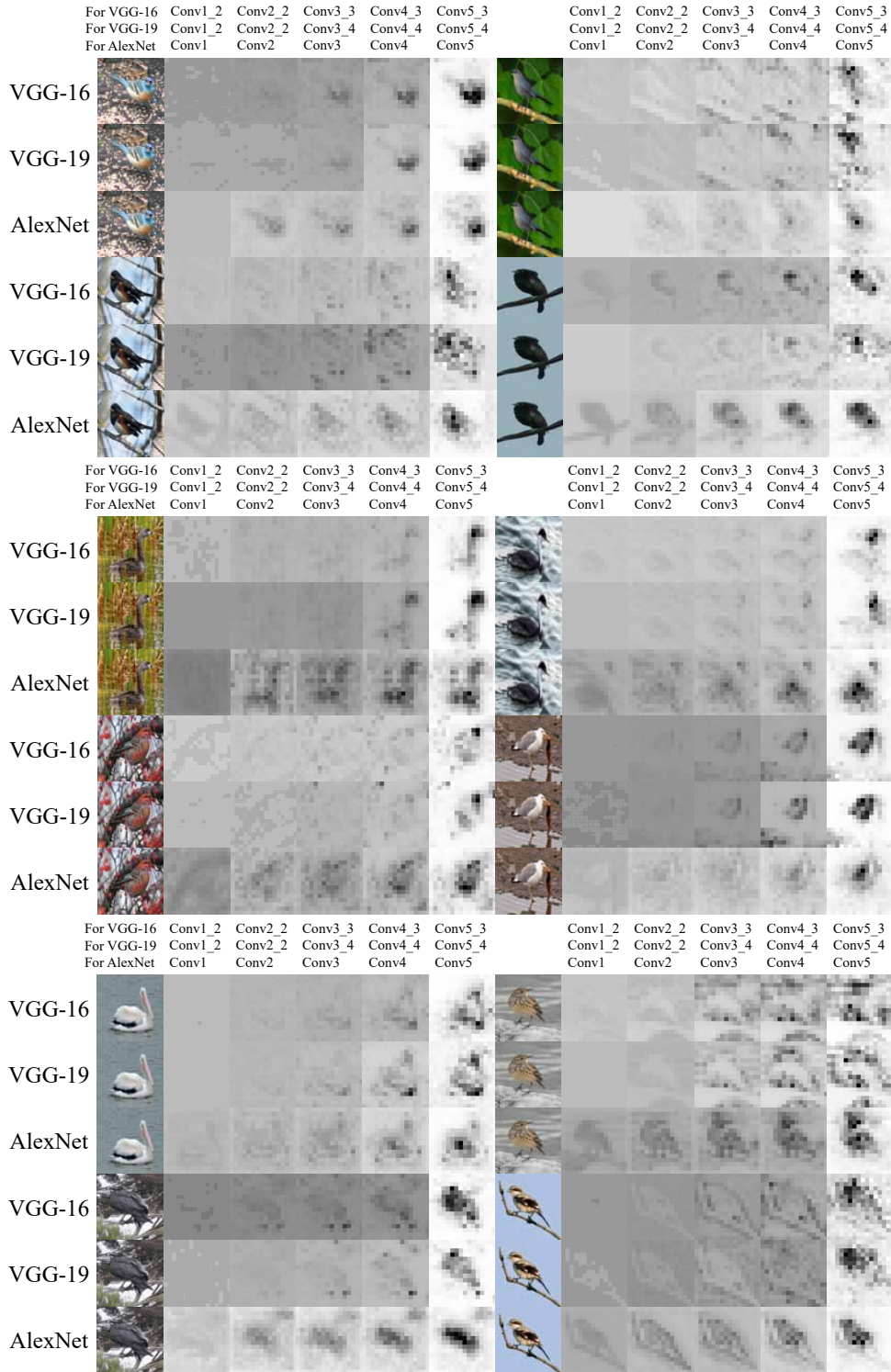
This subsection provides visualization results of CID on the AlexNet learned using the ImageNet dataset. The visualized results can be used to fairly compare the relative importance of the foreground *w.r.t.* the background over different layers.



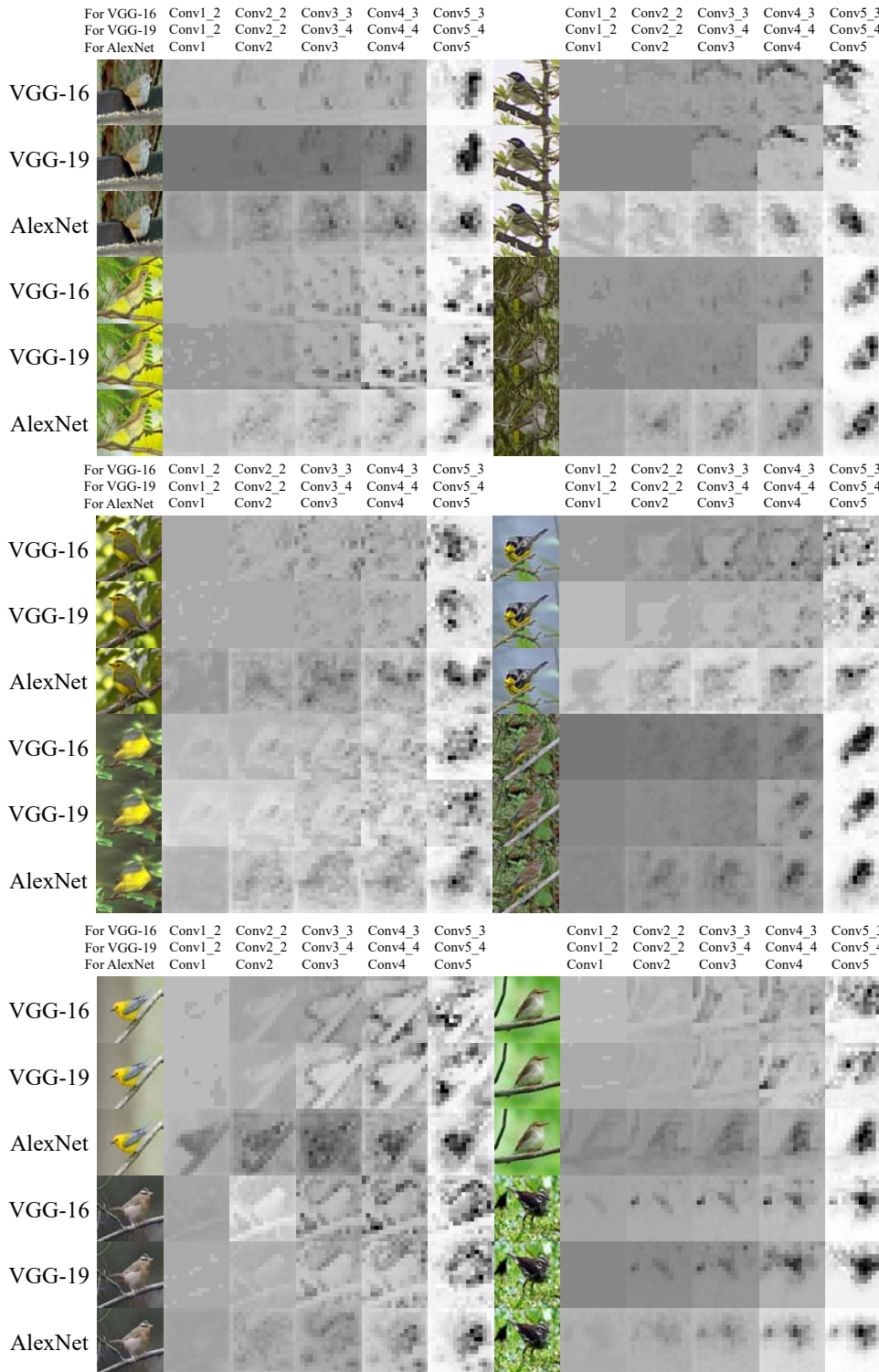
Visualization of CID for AlexNet learned on the ImageNet dataset

G.3. For the AlexNet/VGG-16/VGG-19 learned using the CUB200-2011 dataset

This subsection provides visualization results of CID on the AlexNet/VGG-16/VGG-19 learned using the CUB200-2011 dataset. The visualized results can be used to fairly compare the relative importance of the foreground *w.r.t.* the background over different layers.



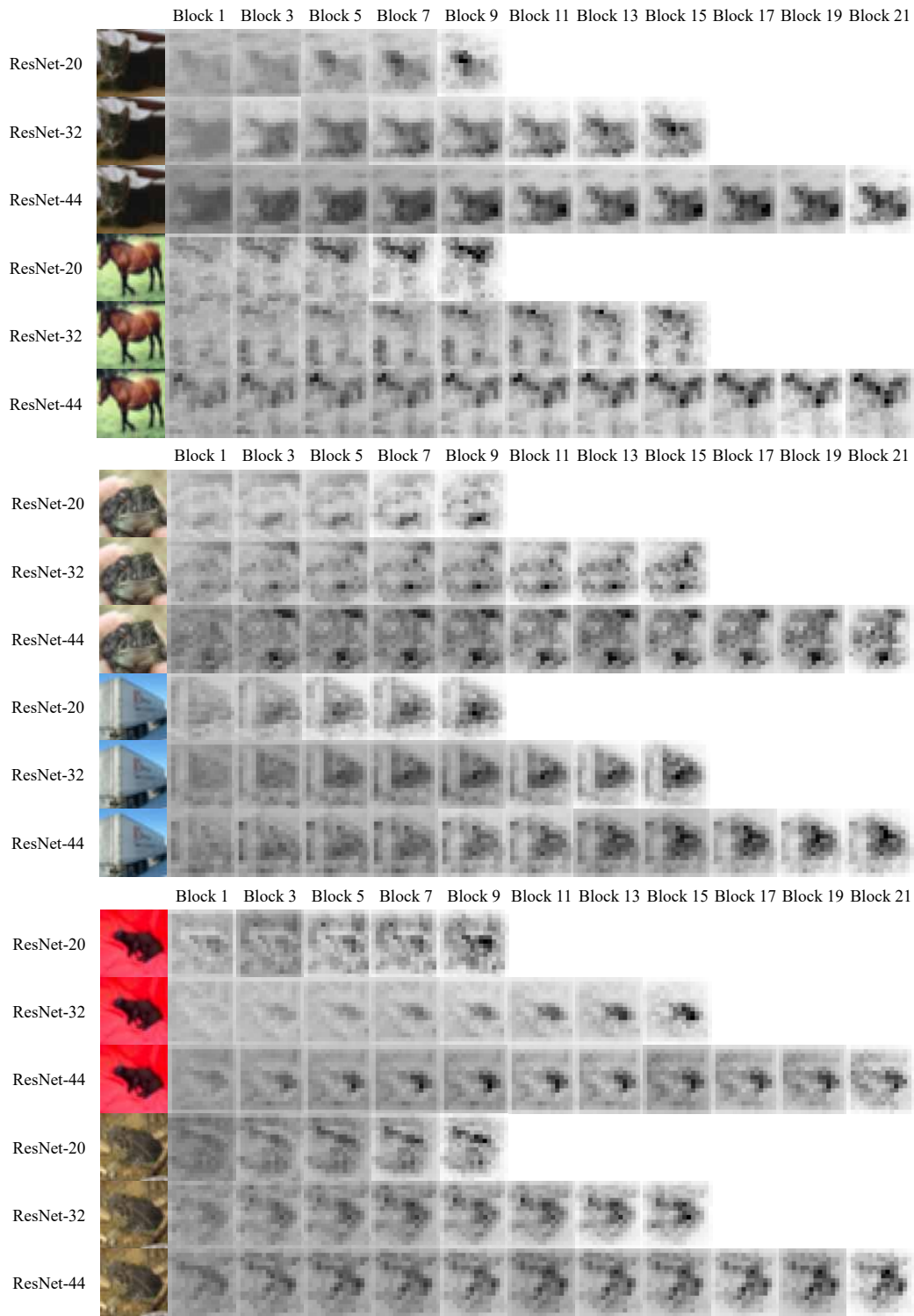
Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding



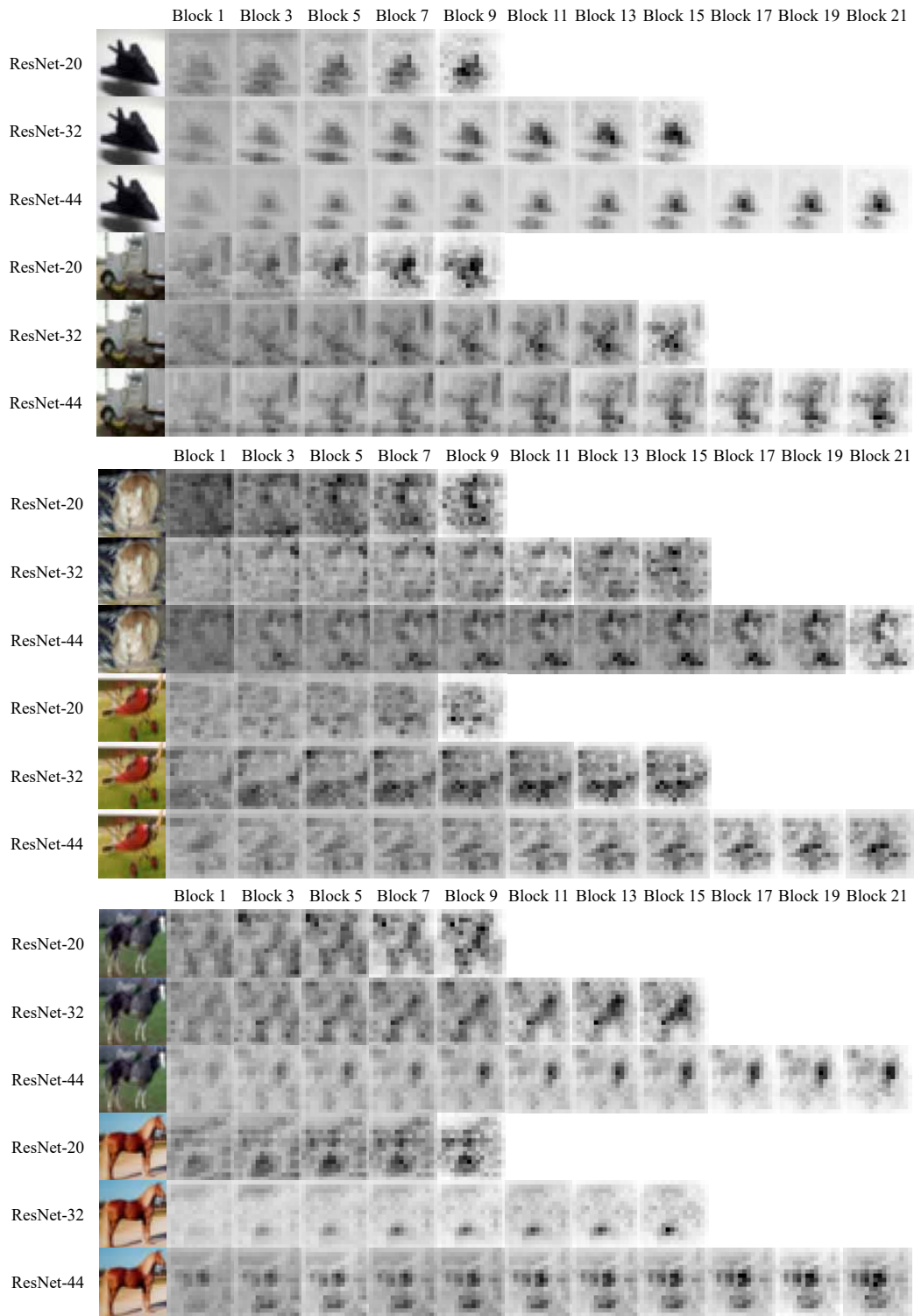
G.4. For the ResNet-20/32/44 learned using the CIFAR-10 dataset

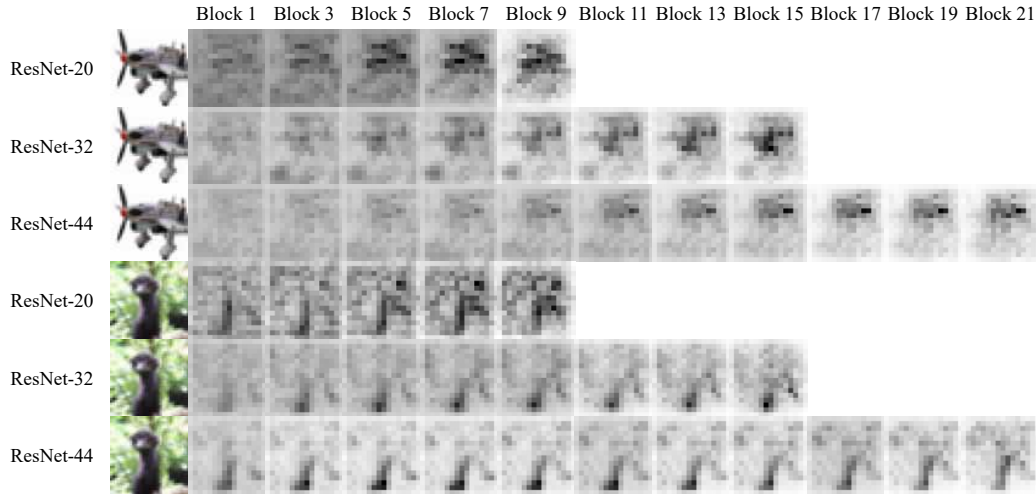
This subsection provides visualization results of CID on the ResNet-20/32/44 learned using the CIFAR-10 dataset. The visualized results can be used to fairly compare the relative importance of the foreground *w.r.t.* the background over different layers.

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



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

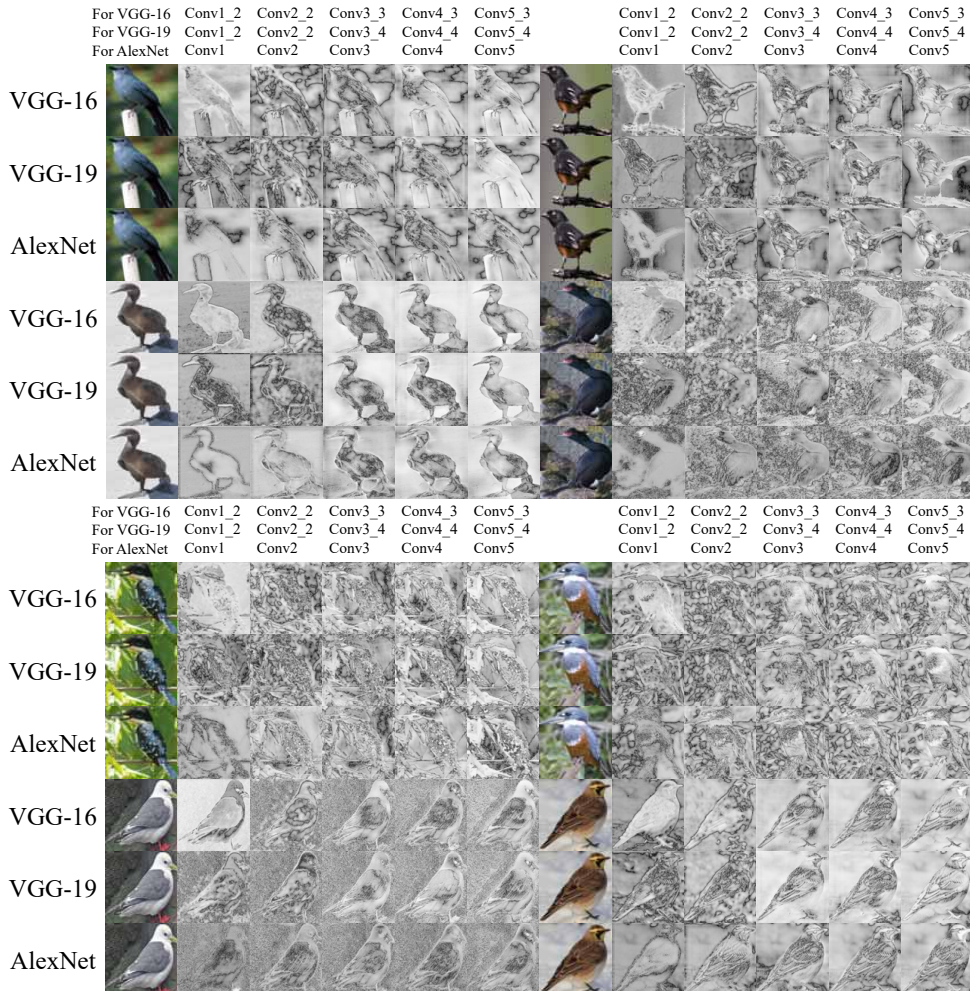




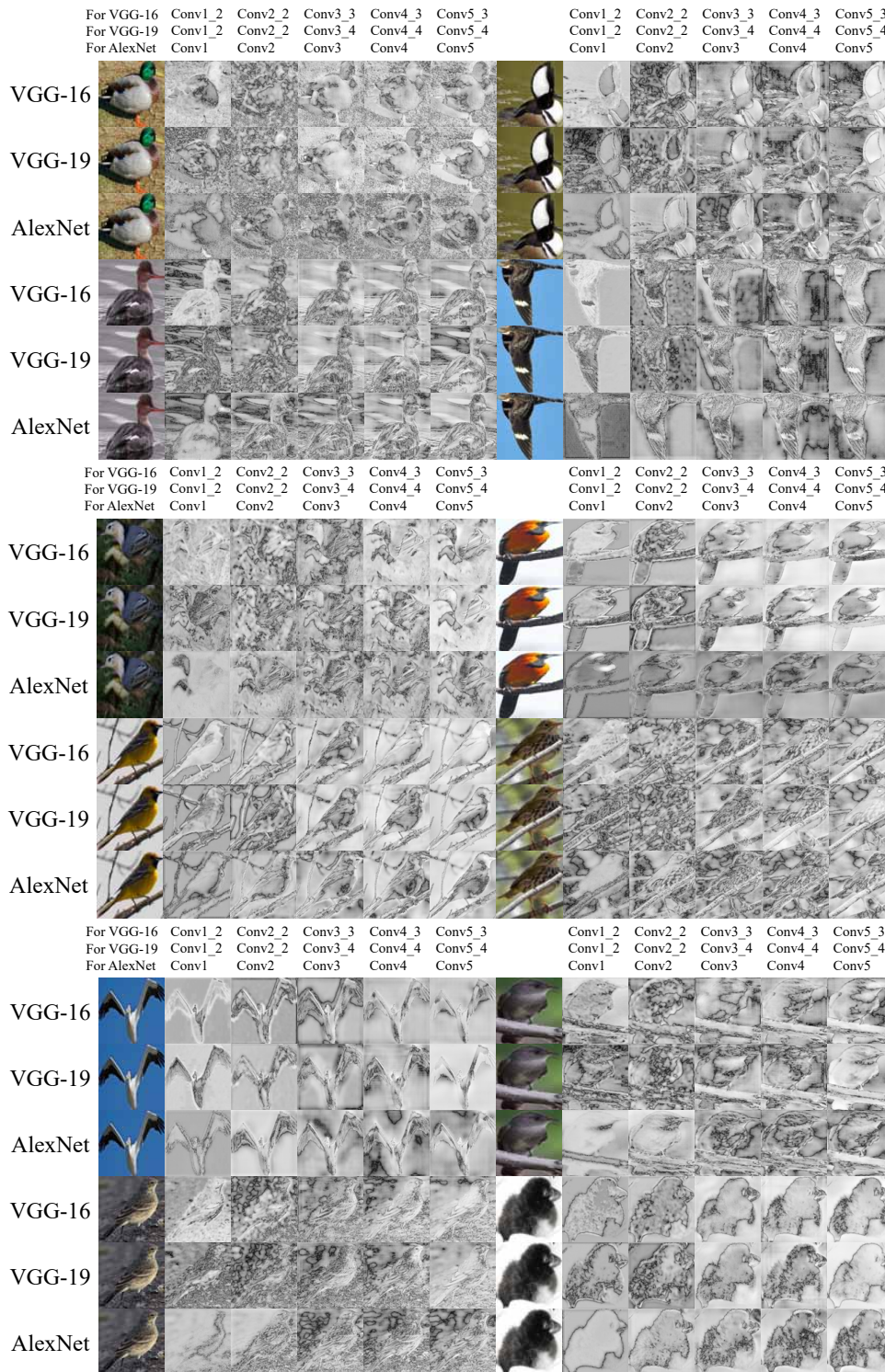
H. Visualization of pixel-wise RU

H.1. For the VGG-16 learned using the CUB200-2011 dataset

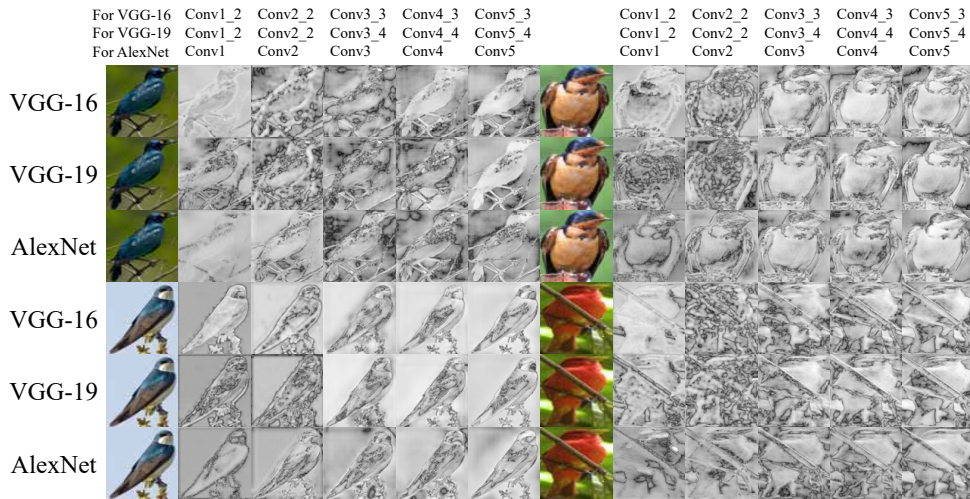
This subsection provides visualization results of RU on the VGG-16 learned using the CUB200-2011 dataset. The visualized results can be used to fairly compare the relative importance of the foreground *w.r.t.* the background over different layers.



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

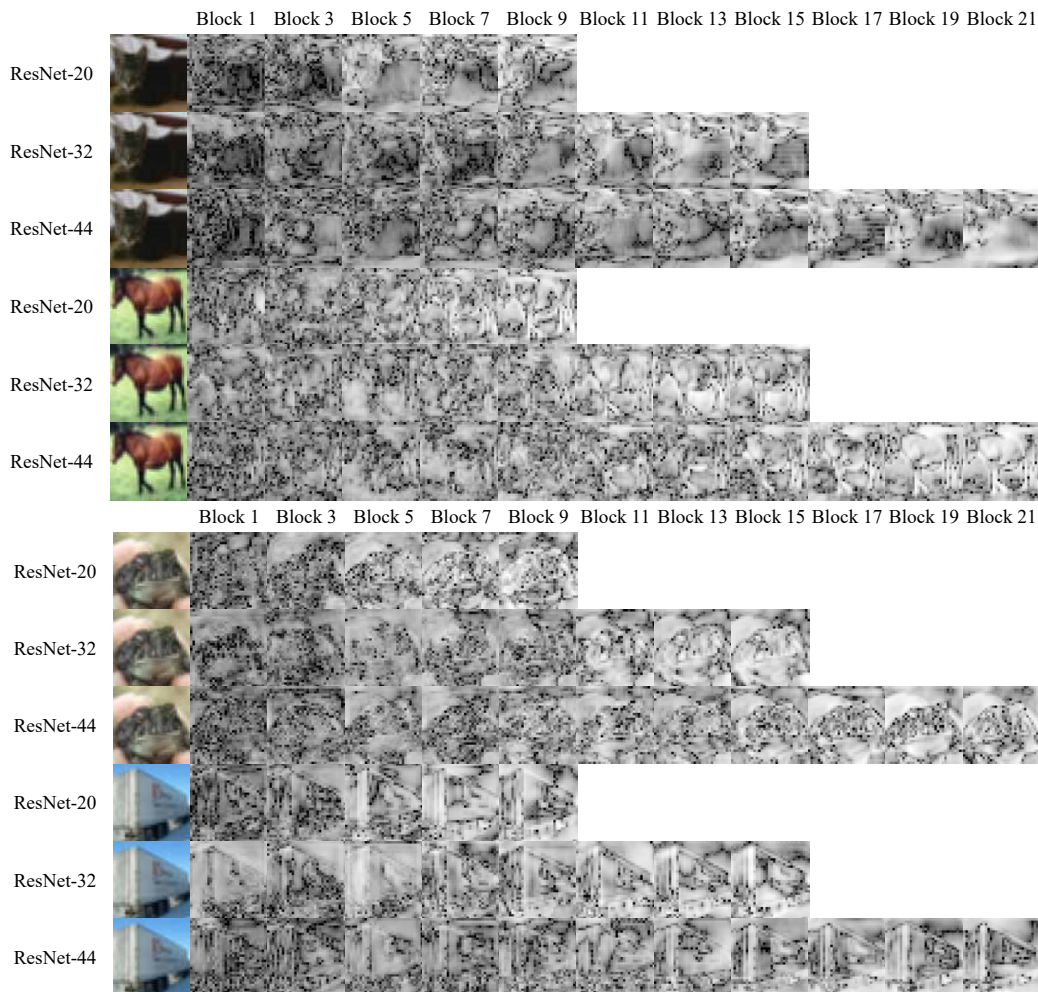


Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding



H.2. For the ResNet-20/32/44 learned using the CIFAR-10 dataset

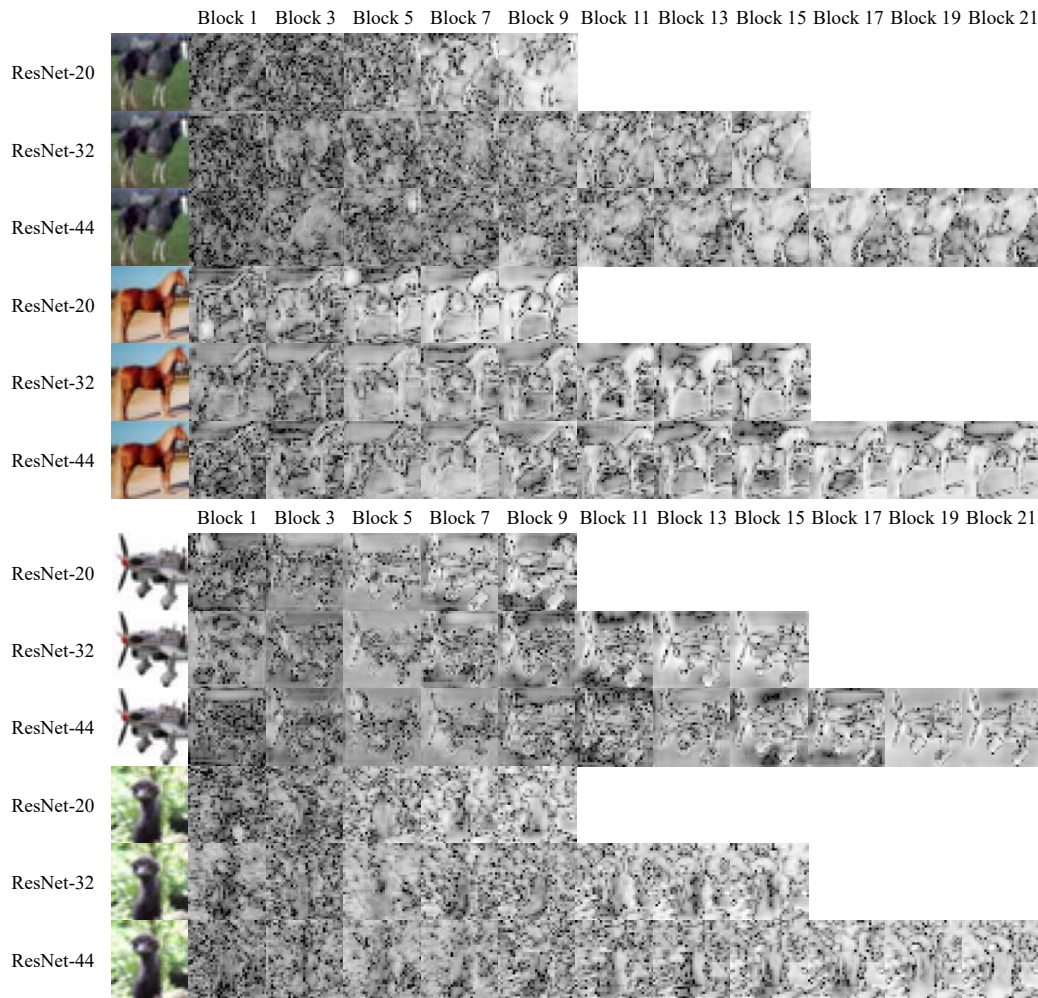
This subsection provides visualization results of RU on the ResNet-20/32/44 learned using the CIFAR-10 dataset. The visualized results can be used to fairly compare the relative importance of the foreground *w.r.t.* the background over different layers.



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

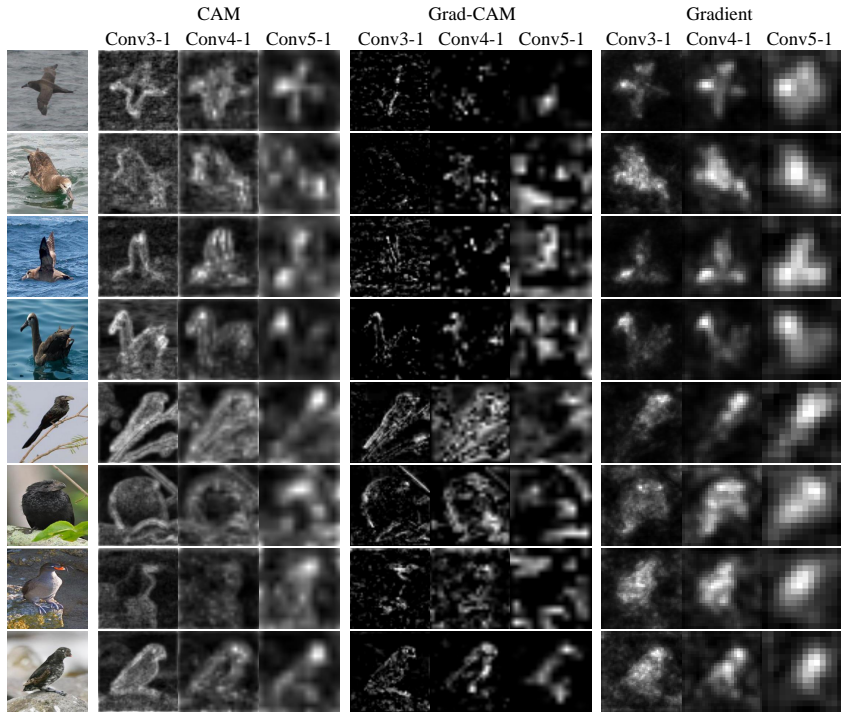


Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding



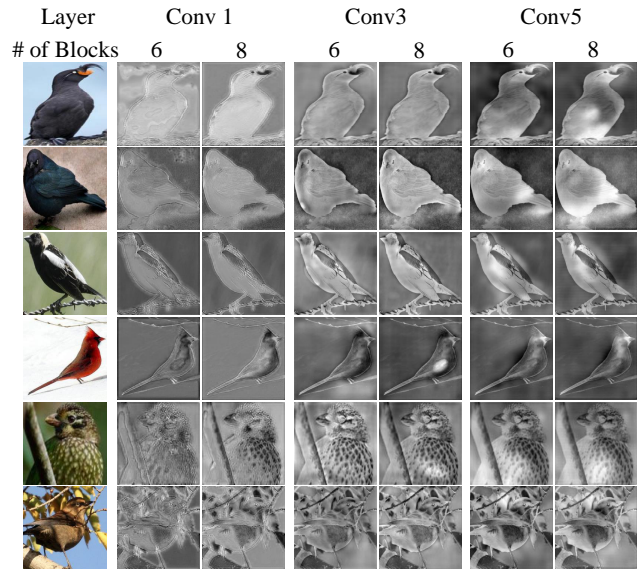
I. Importance maps generated by CAM, Grad-CAM, and Gradient on different layers

The visualization results of importance maps generated by CAM, Grad-CAM, and Gradient on different layers of the VGG-16. The visualized results have been normalized to the unit mean value. According to Table 1 of the paper, CAM, Grad-CAM, and Gradient cannot generate explanation results that enable fair layer-wise comparisons.



J. Comparisons of pixel-wise RU generated by different decoders

In this section, we trained two different decoders for the computation of the metric RU, *i.e.* decoders with six or eight residual blocks, respectively. We visualized RU results in the following figure, which shows that the number of residual blocks in the decoder did not affect the results of pixel-wise RU significantly.

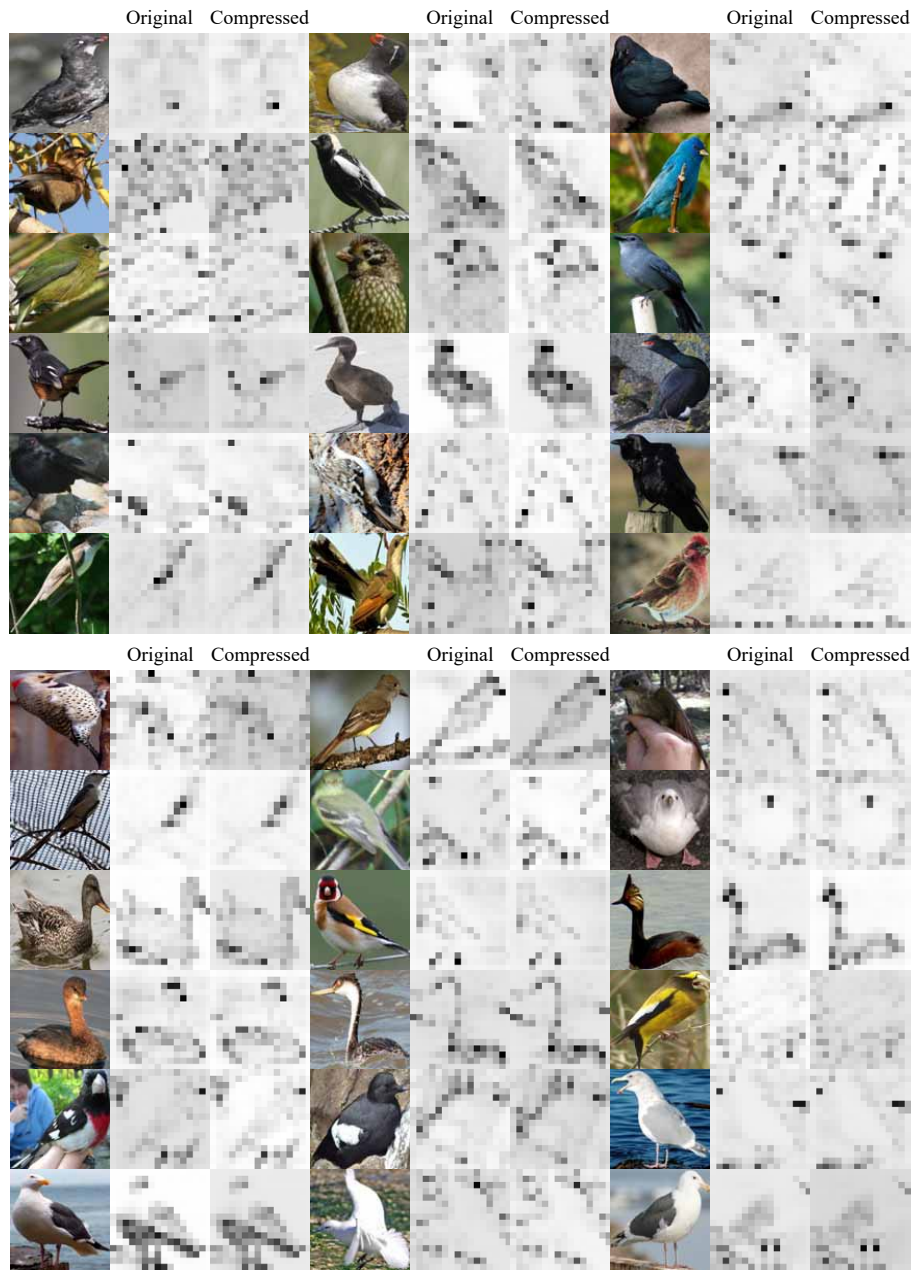


Visualization of RU when RU was computed using different decoders

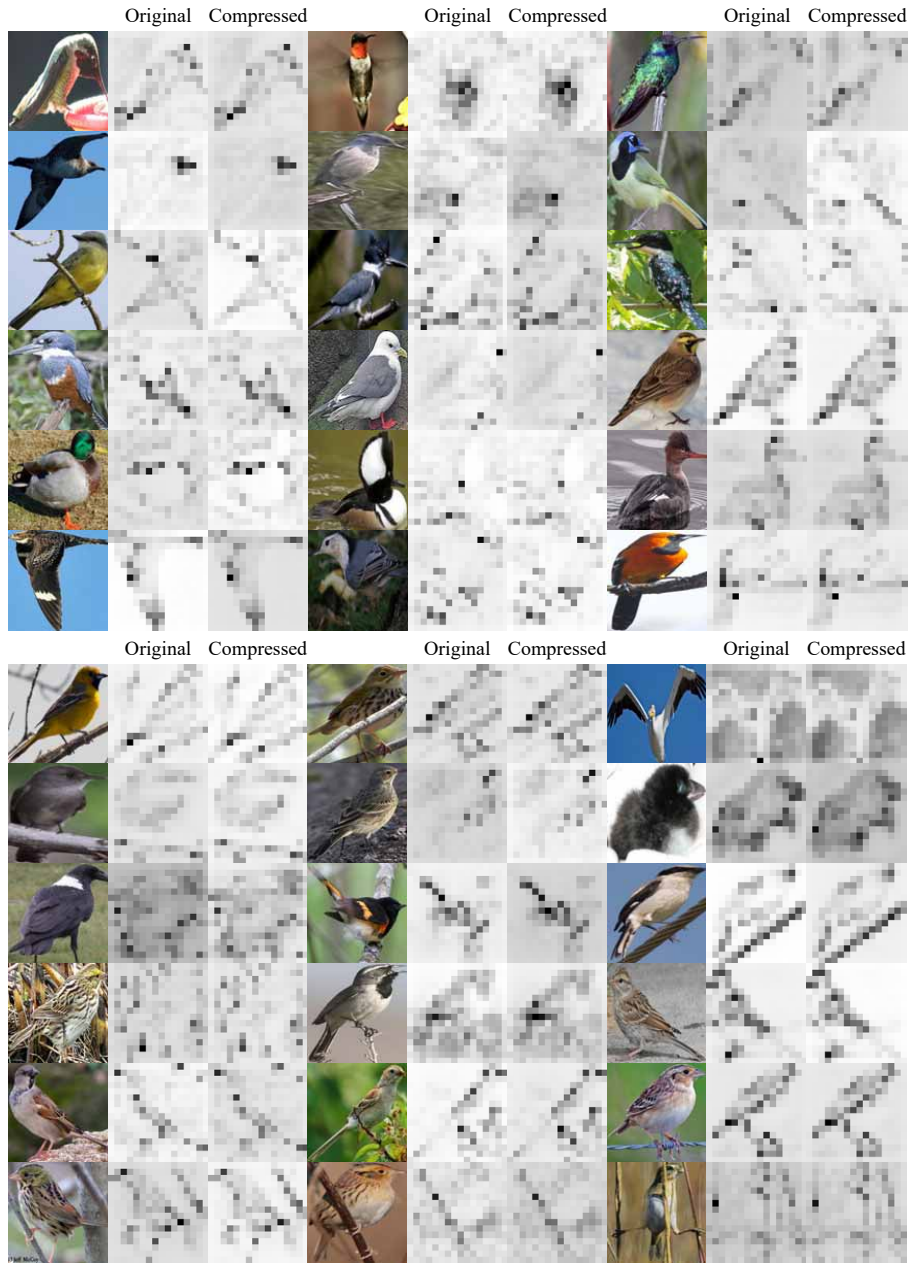
K. Comparisons of pixel-wise CID between the original DNN and the compressed DNN

When we removed 93.3% parameters from the VGG-16 network, the network compression did not significantly change the pixel-wise CID of intermediate-layer features.

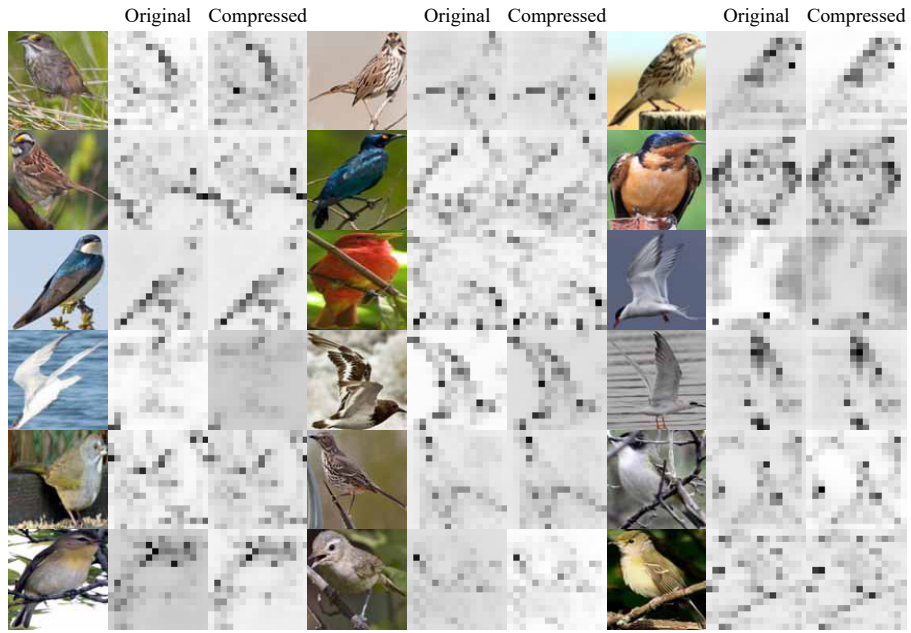
Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding



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

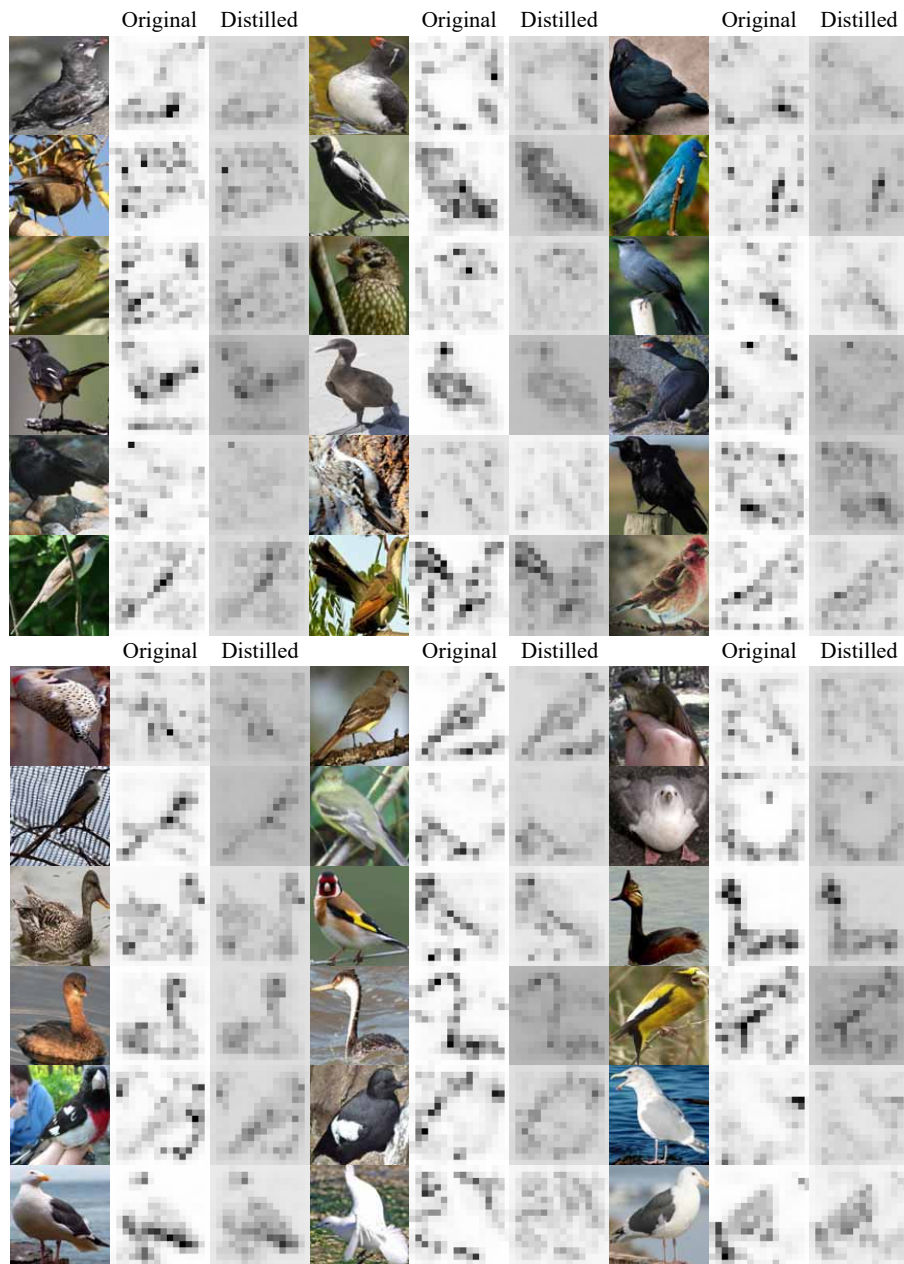


Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding

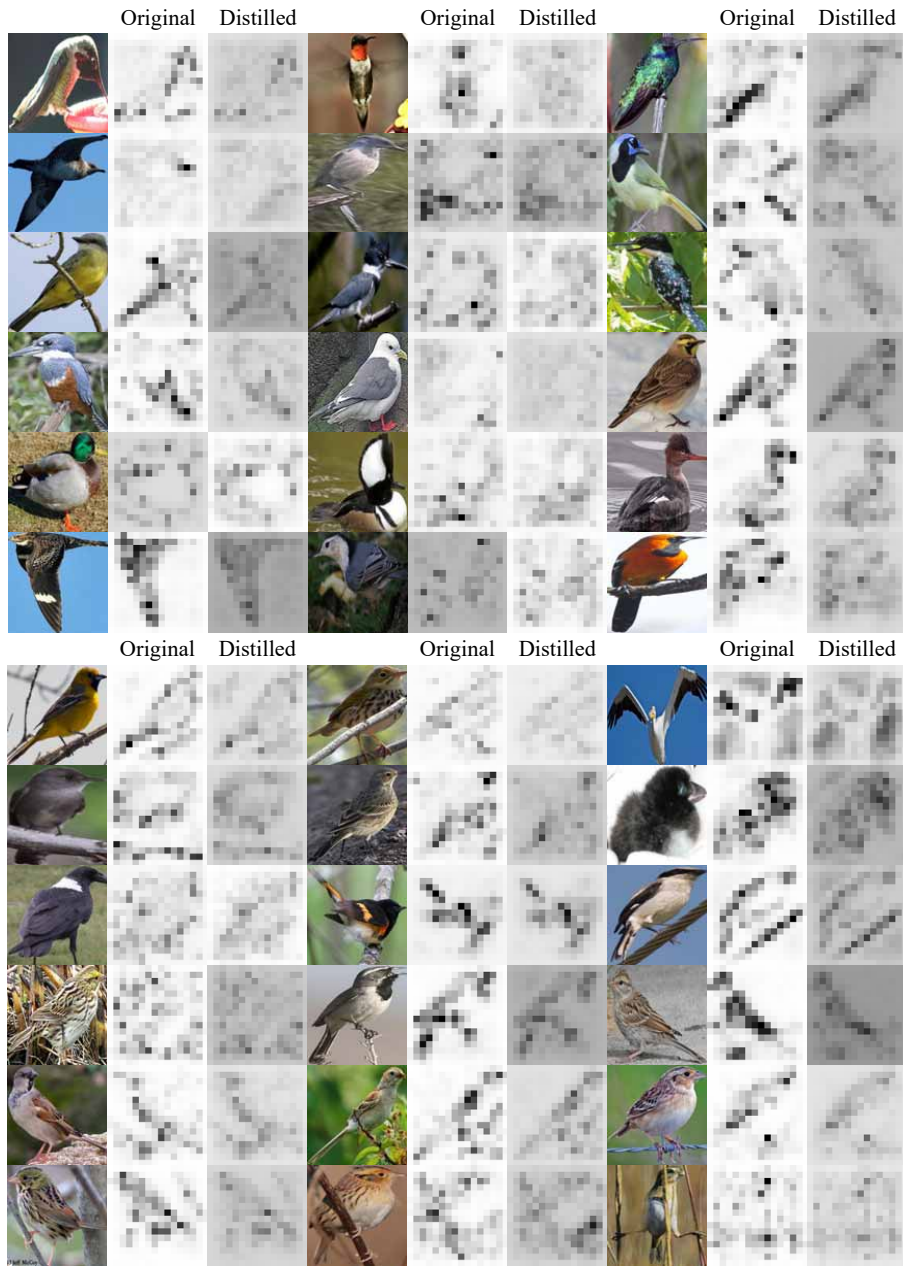


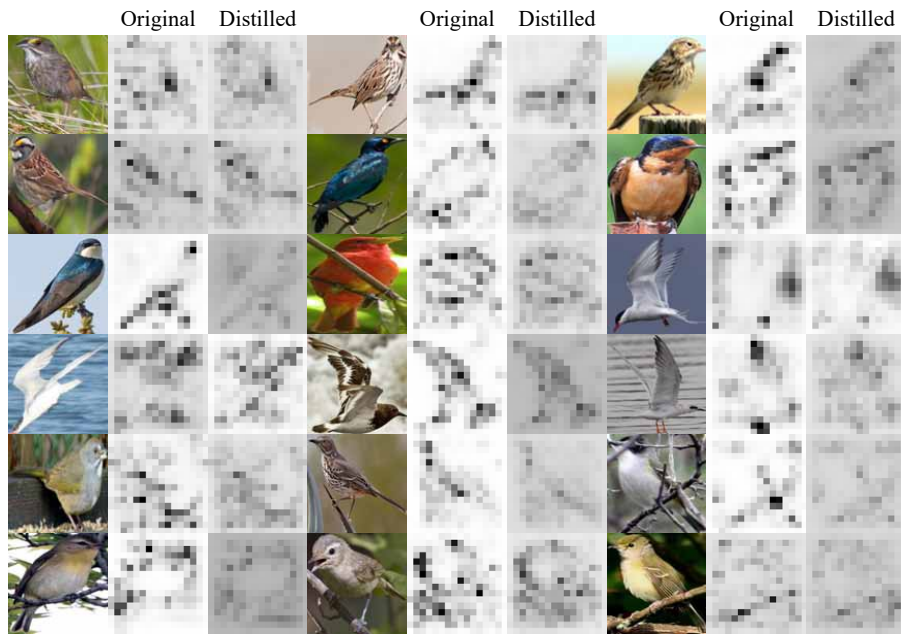
L. Comparisons of pixel-wise CID between the original DNN (the teacher) and the DNN learned via knowledge distillation (the student)

We visualized the pixel-wise CID of VGG-16 networks that were learned using the CUB200-2011 dataset (Wah et al., 2011).



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



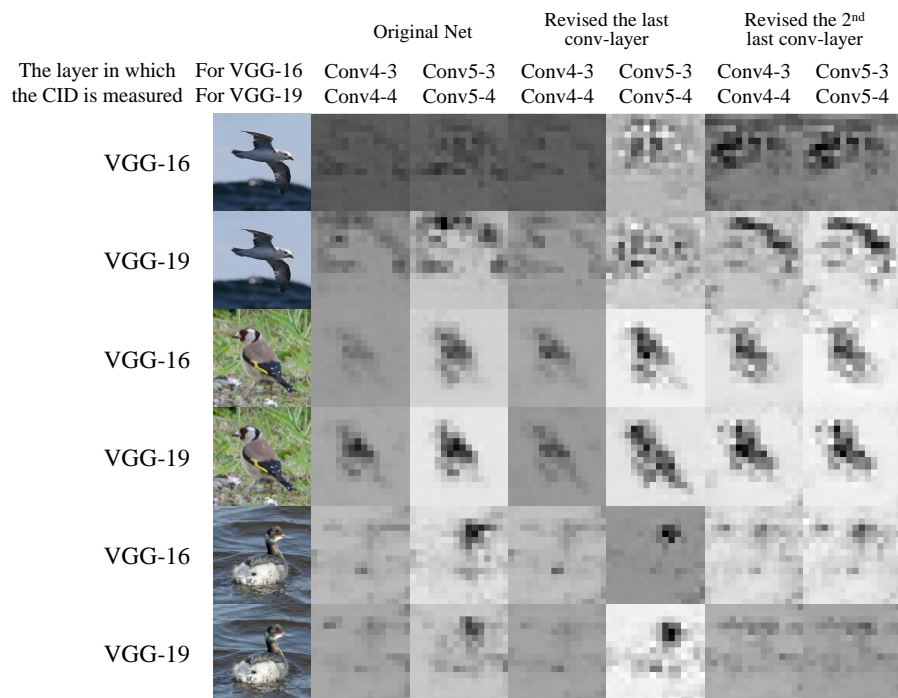
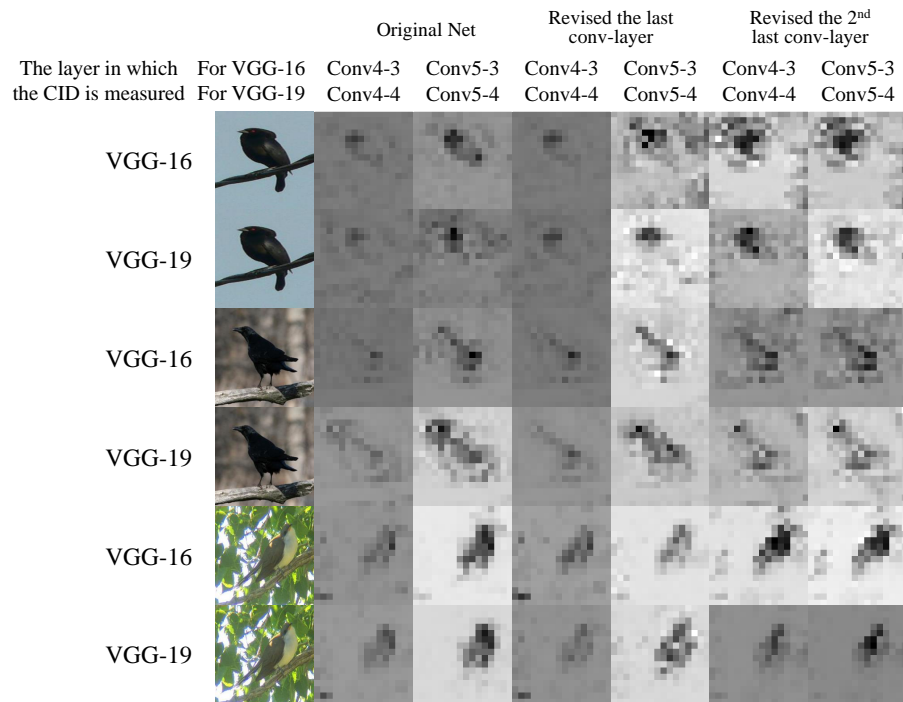


M. Comparisons of pixel-wise CID between the original and the revised DNNs

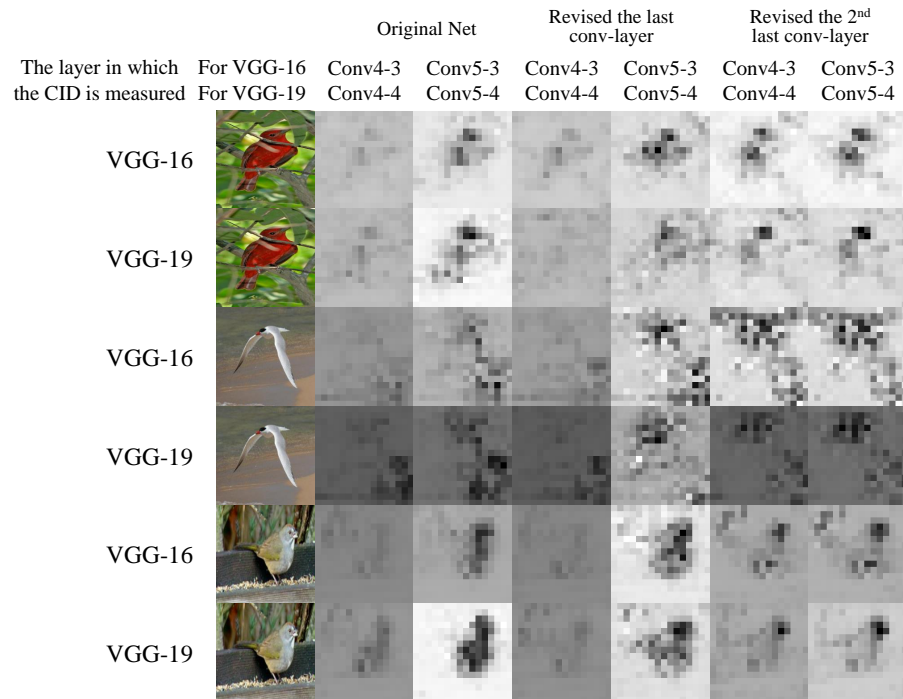
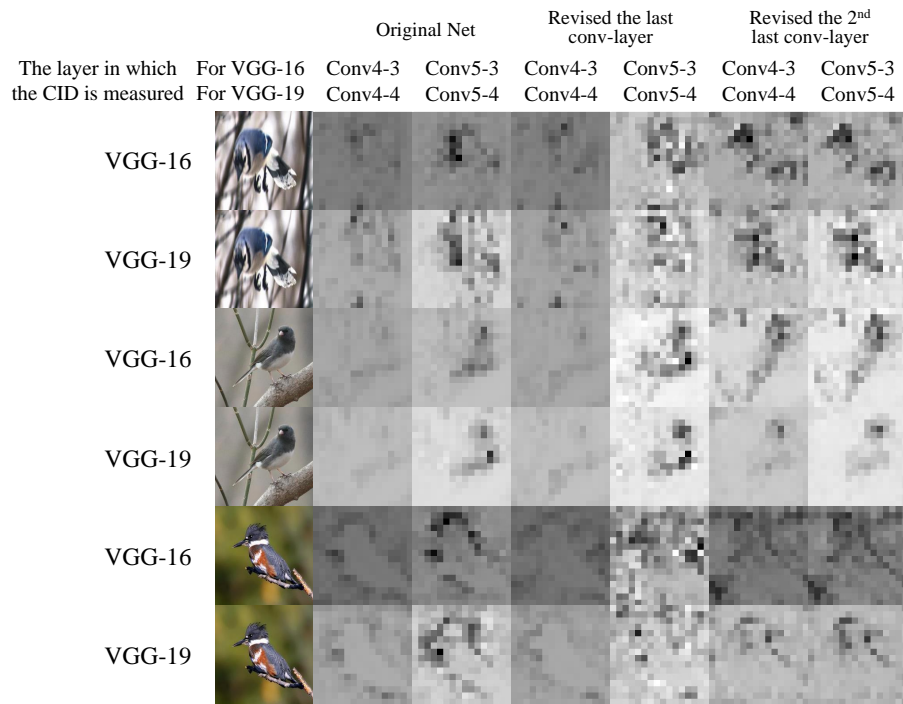
We visualized the pixel-wise CID of the original and damaged networks that were learned using the CUB200-2011 dataset (Wah et al., 2011). We focused on the VGG-16 and VGG-19 networks. For each neural network, we revised either the last convolutional layer or the second last convolutional layer to generate the revised networks.

The layer in which the CID is measured	Original Net	Original Net		Revised the last conv-layer		Revised the 2 nd last conv-layer	
		Conv4-3 Conv4-4	Conv5-3 Conv5-4	Conv4-3 Conv4-4	Conv5-3 Conv5-4	Conv4-3 Conv4-4	Conv5-3 Conv5-4
VGG-16							
VGG-19							
VGG-16							
VGG-19							
VGG-16							
VGG-19							




































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



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



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

The layer in which the CID is measured	For VGG-16 For VGG-19	Original Net		Revised the last conv-layer		Revised the 2 nd last conv-layer	
		Conv4-3 Conv4-4	Conv5-3 Conv5-4	Conv4-3 Conv4-4	Conv5-3 Conv5-4	Conv4-3 Conv4-4	Conv5-3 Conv5-4
VGG-16							
VGG-19							
VGG-16							
VGG-19							
VGG-16							
VGG-19	