CATEGORY DISENTANGLED CONTEXT: TURNING CATEGORY-IRRELEVANT FEATURES INTO TREASURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have achieved great success in computer vision, thanks to their ability in extracting category-relevant semantic features. On the contrary, irrelevant features (e.g., background and confusing parts) are usually considered to be harmful. In this paper, we bring a new perspective on the potential benefits brought by irrelevant features: they could act as references to help identify relevant ones. Therefore, (1) we formulate a novel Category Disentangled Context (CDC) and develop an adversarial deep network to encode it; (2) we investigate utilizing the CDC to improve image classification with the attention mechanism as a bridge. Extensive comparisons on four benchmarks with various backbone networks demonstrate that the CDC could bring remarkable improvements consistently, validating the usefulness of irrelevant features.

1 INTRODUCTION

With the emergence of deep neural networks, their performance on most vision tasks is surpassing human-level. It has reached a consensus that the success of deep networks is brought by their powerful ability in extracting high-level semantic features (Simonyan et al., 2014).

To take more insight into the internal behavior, researchers explain it in the view of attention. The attention is a physiological mechanism which describes the phenomenon that human's perception system could focus on the object of interest (OOI) while suppressing the background. In the last few years, more and more evidences have shown that deep networks originally have such ability to locate OOI (e.g., the category-relevant regions) even without requiring any explicit supervision (Zhou et al., 2015; 2016), therefore enabling them to encode category-relevant features.

To obtain more category-relevant features, researchers design various powerful networks, as stronger networks usually have better attention (Zagoruyko & Komodakis, 2017). However, since the networks are deeper (He et al., 2015a) or wider (Zagoruyko & Komodakis, 2016), the overhead increases accordingly. Other researchers instead control the networks' attention via formulating explicit attention modules, directly refining the networks to encode more category-relevant features. In contrast to the internal guidance, Zagoruyko & Komodakis (2017) attempt to improve the performance of student networks with external guidance by mimicking the attention maps of more powerful teacher networks inspired by knowledge distillation. However, since in both cases the encoded relevant features in the attention maps have a large overlap with those in the backbone networks, the room for improvement is somewhat limited. This brings us to the main topic of this paper: could we solve the problem in the opposite way? More specifically, can we adopt irrelevant features as references and forbid backbone networks to encode them? If so, can backbone networks encode more relevant features with the guidance of pre-extracted category-irrelevant features?

To study these questions, one first needs to specify a proper context that only contains irrelevant features. To that end, we propose to extract a novel Category Disentangled Context (CDC), which is expected to encode all the information of the dataset except that is category-relevant. In this case, the CDC is "complementary" to the category-relevant features. Therefore, the CDC could act as a good reference to help identify the relevant features. We encode the CDC by designing a novel conditional auto-encoder to capture the underlying property of the whole dataset with adversarial training for category disentangling (Mathieu et al., 2016). To demonstrate the potential benefits that could be brought by irrelevant features, we investigate utilizing the CDC to improve the task of image classification. Specifically, we adopt the attention mechanism as a bridge by inferring

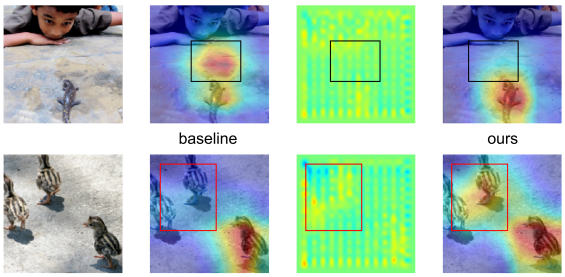

Figure 1: (a) Input images; (b) ResNet-50 originally focuses on some background (black rectangles) and misses part of the targets (red rectangles); (c) by utilizing the attention maps Ms inferred from the CDCs, which indicate category-relevant/irrelevant regions to be encouraged/suppressed; (d) ResNet-50 could be guided to correctly identify the objects of interest in both images. Note: we visualize the focuses of networks following (Zagoruyko & Komodakis, 2017).

the attention map from the CDC and then multiplying it with the backbone networks to refine their attention (e.g., OOI). With the CDC as a reference, backbone networks could purify their encoded features by eliminating irrelevant information that is contained in the CDC (see the suppressed focus marked with black rectangles in the top row of Fig. 1), and explore to encode more relevant features that are not in the CDC (see the added focus marked with red rectangles in the bottom row of Fig. 1), and thus improve the performance. To the best of our knowledge, this is the first work that utilizes category-irrelevant features to improve a vision task.

To summarize, the contributions of this work are as follows:

- We introduce a novel CDC that captures the underlying property of the whole dataset except category-relevant information and develop an adversarial network to obtain it.
- We demonstrate utilizing the CDC to improve image classification in a novel attention manner.
- We validate the effectiveness of the CDC by extensive evaluations with various backbone networks on four public datasets.

2 Category Disentangled Context

Our aim is to design a model that (1) captures the underlying property of the whole dataset; (2) does not have any category-relevant information. We hypothesize that *the information encoded in this model is complementary to category-relevant features*. Therefore, it could act as a good reference to explore more category-relevant features.

We define the latent features $\boldsymbol{T} \in \mathbb{R}^{C\times H\times W}$ encoded in the above model as "*Category Disentangled Context*", where $\boldsymbol{T}$ is an intermediate 3D tensor derived from an input image x. Fig. 2 demonstrates the structure of our CDC Extraction Network, which is adapted from (Lample et al., 2017) with the following key components.

Conditional Auto-encoder. The general architecture of our CDC Extraction Network is a conditional auto-encoder. Compared with (Lample et al., 2017), (1) we choose the latent features encoded in a latter layer, since we require $\boldsymbol{T}$ in a larger spatial resolution, but directly let the original latent features in a high resolution will hinder the auto-encoder to learn a good embedding (Hinton & Salakhutdinov, 2006); (2) we add a skip-connection since it has been validated in previous work (Ronneberger et al., 2015) that skip-connection in the auto-encoder could ease training.

Category Disentangling Branch. To remove category-relevant features, we add a category disentangling branch by extending (Lample et al., 2017), which is originally designed for two-attribute disentangling via attribute flipping and thus could not be directly used in our problem. Specifically, we iteratively train the discriminator D to classify the correct category based on $\boldsymbol{T}$ using the cross entropy loss, and then update the auto-encoder to output a new $\boldsymbol{T}$ to fool D, which is achieved by minimizing the predicted confidence of the correct category,

$$L_{fool}(x, y) = Softmax(V(x))[y], \tag{1}$$

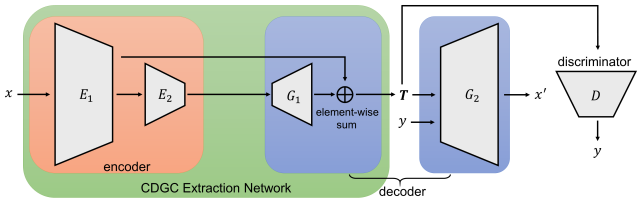

Figure 2: The CDC Extraction Network: given an image-category pair (x, y) as input, the conditional auto-encoder (E_1 and E_2 are encoders, G_1 and G_2 are decoders) outputs a reconstructed image $x^{'}$; D is the discriminator for category disentangling and $\boldsymbol{T}$ is the CDC. Note that only the networks within the green box are executed for generating $\boldsymbol{T}$ in the evaluation stage.

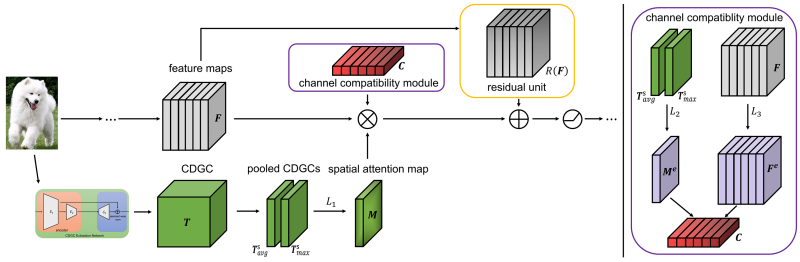

Figure 3: Demonstration of applying the CDC-based framework to the feature maps $\boldsymbol{F}$: L_1 is the network to infer spatial attention map M from the pooled CDCs $[\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]$, L_2 and L_3 are embedding networks to project $[\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]$ and $\boldsymbol{F}$ into the same feature space for calculating channel compatibility; R is the residual unit for adjusting the feature maps after applying attention. Note that the CDC Extraction Network is pre-trained and its parameters are fixed.

where $V(x)$ is the output of the last fully connected layer of the discriminator for image x, and y is the corresponding category. We normalize $V(x)$ using the Softmax function such that each item indicates the predicted probability of one category. We set the weight of L_{fool} as 0.0001.

Repulsive Loss. To guarantee the conditional vector could really work, we add a repulsive loss following (Yu et al., 2018; Wang et al., 2019b) to enforce the discrepancy between images generated with the same $\boldsymbol{T}$ but different conditional vectors to be large enough.

$$L_{repul}(\mathbf{T}) = \max(\delta - ||G_2(\boldsymbol{T}, g(y)) - G_2(\boldsymbol{T}, \mathbf{1} - g(y))||, 0), \tag{2}$$

where $g(y)$ is a function that represents y as a one-hot vector, δ (e.g., 0.01) is the margin to guarantee reasonable changes, and the weight of L_{repul} is 0.001.

3 Utilizing the CDC in Image Classification

To demonstrate the usefulness of the CDC, we investigate utilizing it to improve image classification, which is the basis for almost all the other vision tasks. Specifically, we adopt the attention mechanism as a bridge to utilize the CDC by inferring the attention map from it and then multiplying with the feature maps of backbone networks. Unlike the traditional positive attention mechanism that amplifies category-related neurons (Hu et al., 2018; Woo et al., 2018), our mechanism is in the opposite way, whose goal is to suppress irrelevant neurons. However, unlike in positive ones, where amplified neurons could easily contribute to the network, suppressed neurons will leave activated to deteriorate the performance. Therefore, we multiply the inferred attention maps with the feature maps before activation, and then add a Residual Unit to learn a residual to adjust the suppressed feature maps, such that those irrelevant neurons could be totally inactivated by ReLU. Furthermore, since the inferred attention maps from the CDC are very noisy (see Fig. 4(b)), we adopt a Channel Compatibility Module to alleviate the effects on less related channels. For residual networks, we apply the framework to the feature maps outputted by a residual group. For VGG, we choose the feature maps outputted by the convolutional layer that is just before a maximum pooling layer and have the same resolution as the CDC to apply our framework (see Fig. 3 for demonstration). In the following, we will describe the key components.

Infer Attention Maps. Instead of computing the attention maps from 3D tensors directly (Wang et al., 2017), we conduct two pooling operations to reduce the computational complexity as in (Woo et al., 2018). Specifically, given the CDC $\boldsymbol{T}$ extracted from x, we conduct an average pooling

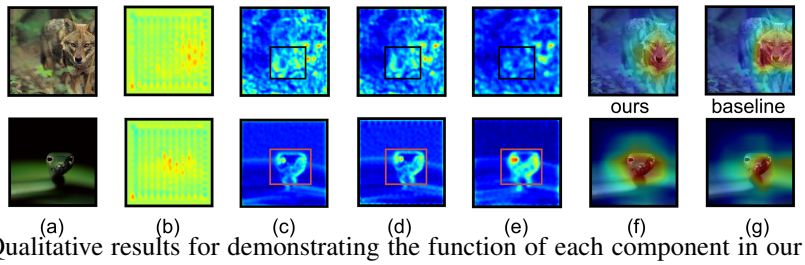

Figure 4: Qualitative results for demonstrating the function of each component in our framework: (a) input images; (b) inferred attention maps Ms from the CDCs; (c) the focused areas by the second residual group of ResNet-50; (d) the focused areas after applying channel-aware suppression/amplification; (e) the focused areas after applying RU; (f) the focused areas by the last residual group; (g) the focused areas by the last residual group of baseline ResNet-50.

and a maximum pooling along the channel axes, generating $\boldsymbol{T}^s_{avg}$ and $\boldsymbol{T}^s_{max}$. Then we infer the attention map $M(\boldsymbol{T}) \in \mathbb{R}^{1\times H\times W}$ (will be abbreviated as M) by forwarding $[\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]$ to the convolutional network L_1 (see Fig. 3):

$$M(\mathbf{T}) = L_1([AvgPool(\boldsymbol{T}); MaxPool(\boldsymbol{T})]) = L_1([\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]), \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation.

Channel Compatibility Module. Suppose $\boldsymbol{F}$ (before activation) is immediate feature maps outputted by the classification network that have the same spatial resolution as M. Since M may not be compatible with all the channels of $\boldsymbol{F}$, therefore instead of applying M to all the channels of $\boldsymbol{F}$ equally, we choose only a part of channels to apply M via computing a channel compatibility measure between $\boldsymbol{T}$ and each channel of $\boldsymbol{F}$. Specifically, we compute the compatibility similar as (Jetley et al., 2018) by first projecting $[\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]$ and $\boldsymbol{F}$ into the same high-dimensional feature space via using the embedding networks L_2 and L_3 respectively, and then conducting dot product between the embedded features $M^e \in \mathbb{R}^{1\times H\times W}$ and $\boldsymbol{F}^e \in \mathbb{R}^{C\times H\times W}$ after squeezing the spatial dimensions, and finally normalizing them with a sigmoid function:

$$C_i = Sigmoid(\langle L_2([\boldsymbol{T}^s_{avg}; \boldsymbol{T}^s_{max}]), L_3(\boldsymbol{F})_i \rangle) = Sigmoid(\langle M^e, \boldsymbol{F}^e_i \rangle), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes dot product after squeezing the spatial dimensions, and subscript i indicates the i-th channel. After that, channel-aware suppression could be realized by applying an element-wise multiplication between each channel of $\boldsymbol{F}$ and the attention map M, weighted with the channel compatibility C, resulting in an irrelevant-suppressed feature $\boldsymbol{F}^{'}$:

$$\boldsymbol{F}^{'}_i = \boldsymbol{F}_i \otimes M \otimes C_i, \tag{5}$$

where $\otimes$ denotes conducting expansion and then element-wise multiplication.

Residual Unit. After suppression, the activation of task-irrelevant neurons will be decreased (e.g., from 0.1 to 0.001), but leave activated (see the black rectangle in Fig. 4). Similarly, the activation of some task-relevant neurons will be relatively enlarged, but since the positive signal is weak, their activations are still very low (e.g., from 0.0001 to 0.001, see the red rectangle in Fig. 4). To resolve these issues, we intentionally add a residual unit R, which is a skip connection as in ResNets (He et al., 2015a), to adaptively learn to adjust $\boldsymbol{F}^{'}$ with $\boldsymbol{F}$ as input (e.g., adjust an irrelevant neuron with value 0.001 to -0.001 and thus is inactivated by ReLU or adjust a relevant neuron with value 0.001 to 0.1). Therefore, the final feature maps are $ReLU(R(\boldsymbol{F}) + \boldsymbol{F}^{'})$, with suppressed neurons inactivated.

Multi-layer Extension. Our framework could be applied to intermediate feature maps outputted by multiple layers with moderate additional computational cost. Specifically, for the feature maps outputted by another layer that have the same spatial resolution as $\boldsymbol{F}$, M^e and M could be directly reused, while for the feature maps with a different resolution, we could down/up-sample M^e and M to make them consistent with the resolution of the target feature maps.

4 Experiments and Discussions

We demonstrate the usefulness of the CDC by taking the task of image classification as an example. Therefore, we exhaustively evaluate the CDC-based classification framework on four benchmarks,

Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100	Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100
VGG13	94.18	74.72	VGG16	93.85	73.78
VGG13 + DM	93.47	73.62	VGG16 + DM	93.16	73.42
VGG13 + GM	94.39	75.48	VGG16 + GM	94.15	75.03
VGG13 + CGM	94.45	75.56	VGG16 + CGM	94.18	74.85
VGG13 + CDCGM	**94.71**	**75.81**	VGG16 + CDCGM	**94.33**	**75.24**

Table 1: Comparison results of different context modeling approaches on the CIFAR datasets.

including CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), ImageNet32×32 (Chrabaszcz et al., 2017) and the full ImageNet (Deng et al., 2009), with various network architectures as backbones, including VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016) and Wide ResNet (Zagoruyko & Komodakis, 2016) (we refer Wide ResNet with depth i and widening factor k as WRN-i-k). We first apply ablation studies on CIFAR datasets due to limited computational resources, and then report more comparisons with the state-of-the-art methods on the other datasets. For the implementation details and more experiments, please refer to the Appendix.

4.1 CIFAR Experiments

Context Modeling Methods. We first investigate whether the CDC is the best context for improving classification. In a multi-category problem with n classes (denote x as the data and $y \in \{1, ..., n\}$ as its label), there are four mathematical models that could be utilized for modeling context:

- *Discriminative model (DM)* attempts to compute $p(\mathbf{y}|\mathbf{x})$, with $p(1|\mathbf{x}) + p(2|\mathbf{x}) + ... + p(n|\mathbf{x}) = 1$. This formulation could be implemented with a classification network.
- *Generative model (GM)* attempts to model $p(\mathbf{x})$ without explicitly considering $\mathbf{y}$. Researchers usually model it as an auto-encoder.
- *Conditional generative model (CGM)* attempts to compute $p(\mathbf{x}|\mathbf{y})$. Generally, researchers adopt the architecture of conditional auto-encoder trained in a semi-supervised manner to model it (Cheung et al., 2014).
- *Category disentangled conditional generative model (CDCGM)* attempts to model $p(\mathbf{x})$, with $p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$, namely the model shouldn't be relevant to $\mathbf{y}$ at all. The readers need to distinguish between CDCGM and GM, although GM does not explicitly consider $\mathbf{y}$, the information of $\mathbf{y}$ is still included, while CDCGM explicitly enforces the model without containing any information of $\mathbf{y}$. Our CDC Extraction Network is a CDCGM.

Although all four models could capture the underlying property of a dataset, their abilities vary a lot. Discriminative models focus on classification boundaries, whereas generative models emphasize the data generation process, and thus generative models could carry richer information (Tu, 2007). In addition, among the three generative models, only CDCGM does not encode any information of $\mathbf{y}$.

We investigate the benefits brought by the context encoded in the above four modeling methods. For DM, we choose WRN-16-10, while for GM and CGM, the networks are simply adapted from our CDC Extraction Network by canceling the introducing of conditional vector and/or removing the category disentangling branch. Comparison results on the CIFAR-10 and CIFAR-100 datasets are reported in Tab. 1. It could be seen that both VGG13 and VGG16 obtain a certain amount of improvements by "attending" the context computed by generative models. In addition, the final results of attending the context in GM and CGM are very similar (e.g., 94.39% and 94.35% on CIFAR-10), since both of their models are relevant to the category, no matter explicitly modeled or not. Obviously, our approach that adopts the CDC performs the best, validating that the CDC could bring more guidance. Interestingly, the performance drops heavily by "attending" the context in a discriminative model, although the approach (Zagoruyko & Komodakis, 2017) that enforces their attention maps to be exactly the same could work (see Tab. 3). The reason is probably that the information of classification boundaries alone is not suitable for the trainable attention mechanism.

Residual Unit (RU). To demonstrate the importance of RU, we train another VGG13/16 with applying the CDC-based framework that has no RU. The results in Tab. 2 show that both networks obtain slight improvement without the help of RU, and the performance is far behind to that of the full framework, validating that the attention maps inferred from the CDC could be hardly utilized without RU. In addition, we also investigate whether the improvement is brought by RU alone via

Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100
VGG13	94.18	74.72
VGG13 + Ours w/o RU	94.24	74.98
VGG13 + Ours w/o CC	94.60	75.62
VGG13 + Ours w/o RL	94.41	75.37
VGG13 + Ours▲	94.40	74.48
VGG13 + Ours▼	94.06	75.20
VGG13 + Ours	**94.71**	**75.81**
VGG13 + Ours after ReLU	94.51	75.33
VGG13 + RU	94.22	75.03

Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100
VGG16	93.85	73.78
VGG16 + Ours w/o RU	94.04	74.21
VGG16 + Ours w/o CC	94.28	74.54
VGG16 + Ours w/o RL	94.19	74.36
VGG16 + Ours▲	94.23	74.73
VGG16 + Ours▼	93.95	74.11
VGG16 + Ours	**94.33**	**75.24**
VGG16 + Ours after ReLU	94.12	74.71
VGG16 + RU	94.06	73.84

Table 2: Ablation studies on the CIFAR-10/100 datasets. Note that Ours▲ indicates the results of applying CDGC to one layer with 16×16 resolution, Ours▼ indicates the results of applying CDGC to one layer with 8×8 resolution, and Ours indicates the results of applying CDGC to both layers.

Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100
VGG13	94.18	74.72
VGG13 + AT (WRN-16-10)	94.43	75.58
VGG13 + Ours	**94.71**	**75.81**
VGG16	93.85	73.78
VGG16 + AT (WRN-16-10)	94.23	74.76
VGG16 + Ours	**94.33**	**75.24**

Architecture	Top-1 Acc(%) CIFAR-10	CIFAR-100
WRN-16-10	94.58	77.72
WRN-16-10 + AT (WRN-28-10)	94.86	78.77
WRN-16-10 + Ours	95.23	79.06
WRN-16-10 + SE	95.81	80.38
WRN-16-10 + SE + Ours	**96.02**	**80.86**
WRN-16-10 + CBAM	95.09	79.51
WRN-16-10 + CBAM + Ours	95.60	80.38
WRN-28-10	95.18	79.15

Table 3: Comparison results on CIFAR-10/100. Note that the architecture described in the bracket after AT denotes the corresponding teacher network.

applying RU to the baseline networks without using the attention mechanism. The limited improvements reported in Tab. 2 validate that RU could not work without the attention framework. Please refer to Fig. 4(d,e) and Fig. 6(d,e) in the Appendix to see the qualitative effects brought by RU.

Repulsive Loss (RL). We demonstrate the importance of RL by training another CDC Extraction Network without adding RL, and then applying the corresponding CDC to VGG13/16 to evaluate the performance. The results in Tab. 2 show that the framework using the CDC extracted without adding repulsive loss could still improve the classification performance, but is consistently worse than that of the full version, validating the importance of repulsive loss.

Channel Compatibility. To investigate the benefits brought by channel compatibility (CC), we compare the results of using CC or not. The results in Tab. 2 show that, without channel compatibility, the models with applying our framework could still outperform the baseline networks, but are worse than their full versions, validating the usefulness of channel compatibility.

Multi-layer. We investigate the usefulness of applying the framework to multiple layers. The results in Tab. 2 show that the performance of only applying the CDC to the layer with 16×16 (8×8)'s output is worse than applying it to both two layers, validating the usefulness of extending to multiple layers. We have also tried to apply the CDC to more layers, but the improvement is very limited compared to the increased overhead.

Apply Attention After ReLU. We investigate whether we could apply the framework to the feature maps after activation. The results in Tab. 2 show that our framework could also work, but the performance is a bit worse than that of applying attention before activation, validating that our framework could utilize ReLU to inactivate irrelevant features instead of only suppressing them.

Comparisons with the State-of-the-art Methods. We make comprehensive comparisons with the state-of-the-art methods using various backbone networks (e.g., VGG and WRN) on the CIFAR datasets. The results in Tab. 3 show that networks with applying the CDC-based framework could bring substantial improvements consistently. We would like to point out that by applying the CDC-based framework, WRN-16-10 achieves comparable results with WRN-28-10, validating that our framework provides a useful alternative to simply deepening or widening networks. Besides, our method outperforms the approach of Attention Transfer (AT) (Zagoruyko & Komodakis, 2017) in all cases including two VGG networks with WRN-16-10 as the teacher and WRN-16-10 with WRN-28-10 as the teacher. Note that, due to the degradation problem (He et al., 2015a), VGG16 performs slightly worse than VGG13, which also indicates the importance of exploring other directions to improve networks. Although the accuracies of Squeeze-and-Excitation Networks (SE) (Hu et al., 2018) and Convolutional Block Attention Module (CBAM) (Woo et al., 2018) are slightly higher than ours, their methods could be further improved (e.g., $\sim$0.5%/0.9% for CBAM) by combining our framework, indicating that our framework is complementary to traditional attention-based methods.

Architecture	Acc(%) Top-1	Top-5
ResNet-32	37.31	62.97
ResNet-32 + AT (ResNet-56)	37.20	62.74
ResNet-32 + Ours	**40.37**	**66.06**
ResNet-32 + SE	37.41	63.02
ResNet-32 + SE + Ours	40.11	65.82
ResNet-32 + CBAM	37.82	63.35
ResNet-32 + CBAM + Ours	40.22	65.88

Architecture	Acc(%) Top-1	Top-5
ResNet-56	42.35	68.01
ResNet-56 + AT (ResNet-110)	42.48	68.18
ResNet-56 + Ours	44.43	70.18
ResNet-56 + SE	42.53	68.29
ResNet-56 + SE + Ours	**45.08**	**70.57**
ResNet-56 + CBAM	43.18	68.92
ResNet-56 + CBAM + Ours	44.55	70.06

Architecture	Acc(%) Top-1	Top-5
ResNet-110	49.08	74.35
ResNet-110 + AT (ResNet-164)	49.03	74.11
ResNet-110 + Ours	**51.36**	**75.93**
ResNet-110 + SE	49.18	74.32
ResNet-110 + SE + Ours	50.29	75.17
ResNet-110 + CBAM	49.43	74.55
ResNet-110 + CBAM + Ours	50.91	75.72

Table 4: Classification results on ImageNet32×32. Note that the architecture described in the bracket after AT denotes the corresponding teacher network.

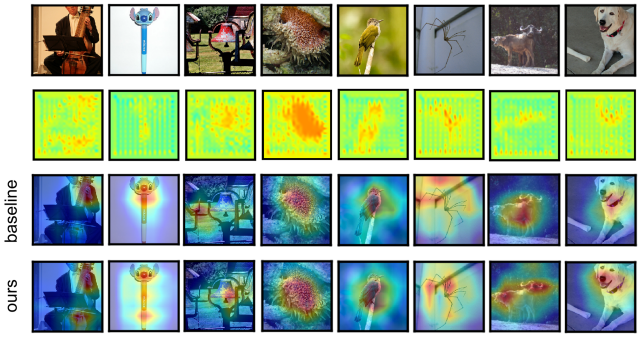

Figure 5: Row 1: input images; Row 2: the inferred attention maps Ms from the CDCs; Row 3: the focused areas by baseline ResNet-50; Row 4: the focused areas by ResNet-50 after applying the CDC-based framework.

4.2 IMAGENET32×32 EXPERIMENTS

To demonstrate the superiority of our method, we perform extensive comparisons with the state-of-the-art methods on ImageNet32×32 (Chrabaszcz et al., 2017). Tab. 4 shows that all three ResNets obtain substantial improvements (e.g., 2∼3% on top-1 accuracy) after applying the CDC-based framework, demonstrating that our approach could generalize well on the large-scale dataset with more categories. In contrast, CBAM- and SE-based approaches improve the baseline networks slightly. Whereas, for the AT-based methods, we could not see any improvement, which probably due to AT-based methods require attention maps to be with high resolutions. In addition, by combining our framework with CBAM or SE, we could see another improvement, validating that utilizing the CDC in an attention manner is complementary to traditional attention-based methods.

Complexity. To make our framework scalable, it must provide an effective trade-off between model complexity and performance. Therefore, we use PARAMs (the number of parameters) and FLOPs (floating-point operations per second) to measure the complexity of the framework. Without considering the network for extracting the CDC, the PARAMs of our framework is 0.034M, while that of ResNet-32, ResNet-56, and ResNet-110 is 0.53M, 0.92M and 1.79M. For the FLOPS, our framework is 3.61M while that of ResNet-32, ResNet-56, and ResNet-110 is 68.19M, 125.51M and 254.46M. We do not report the complexity of the CDC Extraction Network, since it is pretrained and could be reused unlimitedly for different models on the same dataset. Indeed, the CDC Extraction Network requires 0.35M PARAMs and 7.21M FLOPs, which are acceptable, especially for its low FLOPs. Overall, our CDC-based framework is scalable.

4.3 FULL IMAGENET EXPERIMENTS

We conduct experiments on the full ImageNet (Deng et al., 2009) to validate our framework could handle high-resolution images. The results on the validation set reported in Tab. 5 show that by applying the CDC-based framework, all backbone networks obtain large margins of improvement (e.g., 2.4% top-1 accuracy increment for ResNet-50). Besides, our framework with all four ResNets as the backbones outperforms the corresponding SE- and CBAM-based methods. Furthermore, we would like to point out that our approach with ResNet-50 performs better than ResNet-101. Overall, our framework perform well on large-scale datasets in high resolutions.

Visualization. To demonstrate how the CDC bootstrap the backbone networks intuitively, we visualize the focused areas by the last residual group of the original ResNet-50 and those after utilizing the CDC following (Zagoruyko & Komodakis, 2017), together with the attention maps Ms inferred from the CDCs in Fig. 5. It could be seen that ResNet-50 originally fails to locate the objects

Architecture	Acc(%) Top-1	Top-5	Architecture	Acc(%) Top-1	Top-5
ResNet-18	70.33	89.38	ResNet-50	75.49	92.43
ResNet-18 + SE	70.48	89.60	ResNet-50 + SE	76.38	92.89
ResNet-18 + CBAM	70.64	89.83	ResNet-50 + CBAM	77.31	93.44
ResNet-18 + Ours	**71.38**	**90.12**	ResNet-50 + Ours	**77.85**	**93.71**
ResNet-34	73.21	91.32	ResNet-101	76.52	93.04
ResNet-34 + SE	73.76	91.56	ResNet-101 + SE	77.28	93.31
ResNet-34 + CBAM	73.89	91.73	ResNet-101 + CBAM	78.39	94.25
ResNet-34 + Ours	**74.97**	**92.11**	ResNet-101 + Ours	**78.63**	**94.42**

Table 5: Classification results on the full ImageNet dataset.

of interest, and all these locations are correctly identified with the guidance of Ms, validating the usefulness of the CDC. For more visualizations, please refer to Fig. 6 in the Appendix.

5 Related Work

Background Modeling. Most studies in the computer vision community focus on modeling objects in the foreground, leaving background modeling (Bewley & Upcroft, 2017) less investigated. Until very recently, some pioneering work (Zhu et al., 2017; Xiao et al., 2020) have demonstrated that background contains many useful hints for improving image recognition, since it usually has strong semantic correlation with the foreground objects. Compared with them, our work is to model the "semantic" background that is not relevant to the foreground objects at all.

Visual Attention is a basic concept in psychology (Bundesen, 1990). In the field of computer vision, current attention-based methods could be classified into two categories. Post-hoc attention methods analyze the attention mechanism mostly to reason for the task of visual classification (Simonyan et al., 2014; Cao et al., 2015; Zhang et al., 2016; Zhou et al., 2016; Selvaraju et al., 2017). Besides analyzing, Zagoruyko & Komodakis (2017) defined the gradient-based and activation-based attention maps, and improved the student networks by mimicking the attention maps of a more powerful teacher network. Trainable attention methods instead incorporate the extraction of attention and task learning into an end-to-end architecture and are mostly applied to query-based tasks. By projecting the query instance into the same high-dimensional space as the target, relevant spatial regions of the query will be highlighted to guide the desired inference (Jetley et al., 2018; Anderson et al., 2018; Chen et al., 2017; Bahdanau et al., 2014). In contrast to those approaches that adopt query-based attention, self-attention based methods attempt to learn the attention maps by themselves (Hu et al., 2018; Wang et al., 2017; Woo et al., 2018; Wang et al., 2019a; Bello et al., 2019; Parmar et al., 2019; Zhao et al., 2020). Our approach consists both the post-hoc and trainable attention modules. For the extraction of the CDC, it could be considered as a post-hoc attention method. The most similar work to ours is (Zagoruyko & Komodakis, 2017), as they also attempt to adopt the "context" of a teacher network to guide the student network. However, the difference is still two-fold. Firstly, they apply the context as a "hard" constraint, enforcing the two attention maps to be exactly the same, such that some valuable information of the student network is forced to be discarded; while we apply it as soft guidance, and thus could take advantage of both networks. Secondly, unlike their approach that obtains the context via a deep network which is a discriminative model, the CDC is encoded in a category disentangled conditional generative model, and thus could bring more guidance. For "attending" the CDC, it is a trainable attention method. Compared with (Hu et al., 2018; Woo et al., 2018) that encourage feature activation on category-relevant regions, our mechanism is in the opposite way. Besides, we adopt the CDC as external knowledge to infer the attention maps instead of totally by the backbone networks themselves. Last but not least, since our key idea is to utilize category-irrelevant features, it is complementary to traditional attention-based methods.

6 Conclusion

We have presented a novel Category Disentangled Context (CDC), which is a kind of category-irrelevant features capturing the underlying property of the whole dataset. We demonstrate its usefulness by utilizing it as a reference to guide image classification networks, with the attention mechanism as a bridge. Extensive experimental results validate the CDC could bring substantial improvements for various backbone networks and is superior to the state-of-the-art methods. In the future, we plan to apply the CDC to handle more complex vision applications, e.g., generating better region proposals and making more accurate predictions on them in object detection. Overall, we think that our interesting findings will help further advance the research on irrelevant features, and understanding convolutional neural networks in general.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pp. 6077–6086, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Irwan Bello, Barret Zoph, Quoc V Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *ICCV*, pp. 3286–3295, 2019.

Alex Bewley and Ben Upcroft. Background appearance modeling with applications to visual object detection in an open-pit mine. *Journal of Field Robotics*, 34(1):53–73, 2017.

Claus Bundesen. A theory of visual attention. *Psychological Review*, 97(4):523, 1990.

Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, pp. 2956–2964, 2015.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pp. 6298–6306, 2017.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2015a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ICCV*, pp. 1026–1034, 2015b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pp. 630–645, 2016.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pp. 7132–7141, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. In *ICLR*, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *NeurIPS*, pp. 5967–5976, 2017.

Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *CVPR*, pp. 3901–3910, 2017.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NeurIPS*, pp. 5040–5048, 2016.

Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, pp. 68–80, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8026–8037, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Zhuowen Tu. Learning generative models via discriminative approaches. In *CVPR*, pp. 1–8, 2007.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pp. 3156–3164, 2017.

Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuanchuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N Metaxas. Sharpen focus: Learning with attention separability and consistency. In *ICCV*, pp. 512–521, 2019a.

Wei Wang, Yuan Sun, and Saman K Halgamuge. Improving mmd-gan training with repulsive loss function. In *ICLR*, 2019b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pp. 3–19, 2018.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv e-prints*, 2020.

Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. PU-Net: Point cloud upsampling network. In *CVPR*, pp. 2790–2799, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff, and Sarah Adel Bargal. Top-down neural attention by excitation backprop. In *ECCV*, pp. 543–559, 2016.

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *ICLR*, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pp. 2921–2929, 2016.

Zhuotun Zhu, Lingxi Xie, and Alan L. Yuille. Object recognition with and without objects. In *IJCAI*, pp. 3609–3615, 2017.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

The CDC Extraction Network.

We adapt our CDC Extraction Network from (Lample et al., 2017). Denote $c(i, j, k)$ as a convolution layer, where each convolution uses kernel of size $i \times i$, with a stride of j, and a padding of k, and let $C(i, j, k)$ be a group of layers with $c(i, j, k)$, BatchNorm (BN) and ReLU. For images with the resolution of 32×32, our encoder consists of the following six groups: $C(4, 2, 1) - C(3, 1, 1) - C(4, 2, 1) - C(3, 1, 1) - C(4, 2, 1) - C(4, 2, 1)$, with the first four groups belonging to E_1. The decoder is symmetric to the encoder, except replacing the convolutions in $C(4, 2, 1)$ with transposed convolutions and appending the conditional vector as additional constant input channels for all the layers of G_2 (see Fig. 2). Therefore, the final resolution of the CDC is 8×8. For images with the resolution of 224×224, the second $C(3, 1, 1)$ group is replaced by $C(4, 2, 1)$, resulting in 28×28's CDC. For images on ImageNet-150K with the resolution of 128×128, the encoder is defined as $C(4, 2, 1) - C(4, 2, 1) - C(4, 2, 1) - C(4, 2, 1) - C(4, 2, 1) - C(4, 2, 1)$ for a 8×8 CDC, and we replace the fourth convolution group with $C(3, 1, 1)$ for a 16×16 CDC. Note that, for the cases with more than ten categories, we additionally adopt a fully connected network with two layers to embed the one-hot category vector into a ten-dimensional conditional vector. We train the CDC Extraction Network in the same way as (Lample et al., 2017). Note that, for each dataset, the CDC Extraction Network is required to be trained only once.

The CDC-based Framework.

R and L_3 are with the same structure of BN-ReLU-$c(1, 1, 0)$-BN-ReLU-$c(3, 1, 1)$-BN-ReLU-$c(1, 1, 0)$, which uses the bottleneck structure to reduce complexities. For the experiments with 32×32 images, channel numbers are reduced to $\frac{1}{2}$ of the original ones, while $\frac{1}{4}$ for 224×224 and 128×128 images. L_1 is with $c(3, 1, 1)$-BN-ReLU-$c(3, 1, 1)$-BN-Sigmoid, and L_2 is with $c(3, 1, 1)$-BN-ReLU-$c(3, 1, 1)$. To fit the CDC Extraction Network, we normalize all the input images into [-1, 1], except for those experiments on the full ImageNet, where mean/std normalization is applied, and thus we should transform them back into [-1,1] before feeding to the CDC Extraction Network. We adopt the two-layer CDC-based framework by adding a 2×2 up-sampling of the CDC.

Backbones.

For VGG, we add BatchNorm (Ioffe & Szegedy, 2015) while leaving Dropout (Srivastava et al., 2014) removed, and use one fully connected layer. The implementation of ResNets and Wide ResNets are identical to the original papers.

Training Details.

We adopt SGD using default parameters as the optimizer with a batch size of 128 for CIFAR datasets and ImageNet32×32, 32 for ImageNet-150K, and 256 for ImageNet. The initial learning rate for CIFAR datasets is set as 0.1, and it is divided by 5 at 60, 120 and 160 epochs for total 200 epochs. For ImageNet32×32 and ImageNet-150K, we start the learning rate with 0.01, and divide it by 5 every 10 epochs for total 40 epochs. For ImageNet, the initial learning rate of ResNets is set as 0.1, and we divide it by 10 at 30, 60 and 90 epochs for a total of 100 epochs. For data augmentation, we follow (Zagoruyko & Komodakis, 2016) on CIFAR datasets and ImageNet32×32 but without mean/std normalization. Similarly, we horizontally flip each image and randomly take a 128×128 crop with 16 pixels padded on each side on ImageNet-150K. While for ImageNet, we apply minimal data augmentation including random resized crop, flip and the simple mean/std normalization. We initialize the weights following (He et al., 2015b) and warm up the training with a learning rate of 0.001 for 2 epochs. The final results are averaged based on three independent runs. For each run, we evaluate the single-crop performance at each epoch on NVIDIA Tesla V100 GPUs and report the best one.

All the CDC Extraction Networks in our experiments are trained with the Adam optimizer, using a learning rate of 0.0002, β_1=0.5, and β_2=0.999. We train the models with a batch size of 32 for 600 epochs on CIFAR datasets, 300 epochs on ImageNet-150K, and 50 epochs on ImageNet32×32 and the full ImageNet datasets.

We reproduce all the networks in PyTorch (Paszke et al., 2019) without any model pre-training.

Table 6: Classification results on ImageNet-150K.

Architecture	Acc(%) Top-1	Top-5
ResNet-101	78.70	94.55
ResNet-101 + AT (teacher: ResNet-152)	78.40	94.35
ResNet-101 + Ours (8×8)	79.00	95.10
ResNet-101 + Ours (16×16)	**80.85**	95.15
ResNet-101 + SE	79.30	94.35
ResNet-101 + SE + Ours	80.65	94.80
ResNet-101 + CBAM	78.80	94.20
ResNet-101 + CBAM + Ours	80.35	**95.50**

A.2 IMAGENET-150K EXPERIMENTS

We also conduct experiments on the ImageNet-150K dataset (Liu et al., 2017), which is a subset of ImageNet and also contains 1k categories but with 150 images per category (148 for training, 2 for testing). We resize all the images into 128×128.

Experimental results are depicted in Tab. 6. It is shown that ResNet-101 obtains the largest improvement by applying the CDC-based framework, compared with three other state-of-the-art methods: CBAM (Woo et al., 2018), SE (Hu et al., 2018) and AT (Zagoruyko & Komodakis, 2017) (e.g., 80.85% vs. 78.40%, 79.30% and 78.80% top-1 accuracy), validating that our approach could generalize well on the datasets with higher resolutions. Particularly, AT could hardly work on ResNet-101 with ResNet-152 as the teacher. This is probably due to ResNet-101 has reached the performance limit in this dataset as with limited training images. Without depending on other networks, CBAM-, SE-based methods and our framework do not suffer from it. In addition, by combining our framework with CBAM or SE, we could see another improvement, validating that utilizing the CDC in an attention manner is complementary to traditional attention-based methods.

Different Resolutions of the CDC. We also investigate whether different resolutions of the CDC will influence the final results. It is shown that the framework with the CDC in a higher resolution (e.g., 16×16) performs better. Therefore, we attempt to obtain higher resolutions of the CDC to make more comparisons. However, we find the CDC Extraction Network could hardly converge, indicating the challenging of disentangling latent features with a high resolution.

A.3 MORE VISUALIZATIONS AND QUALITATIVE RESULTS

Fig. 6 visualizes the focused areas by the last residual group of the original ResNet-50 and the one after applying our proposed CDC-based framework following (Zagoruyko & Komodakis, 2017), together with their corresponding attention maps inferred from the CDCs and some intermediate results to demonstrate the usefulness of each component in our framework.

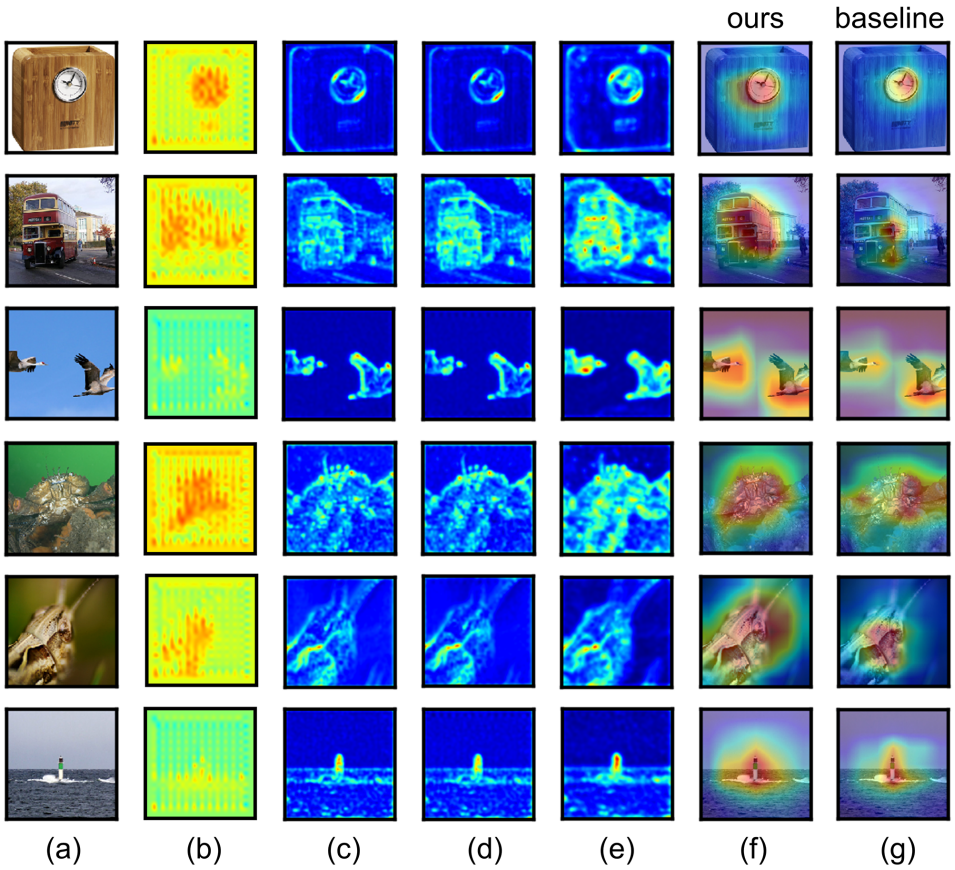

Figure 6: Qualitative results for demonstrating the function of each component in our framework: (a) input images; (b) inferred attention maps Ms from the CDCs; (c) the focused areas by the second residual group of ResNet-50; (d) the focused areas after applying channel-aware suppression/amplification; (e) the focused areas after applying RU; (f) the focused areas by the last residual group; (g) the focused areas by the last residual group of baseline ResNet-50.