

# Distilling Localization for Self-Supervised Representation Learning

Nanxuan Zhao<sup>1\*</sup>

Zhirong Wu<sup>2\*</sup>

Rynson W.H. Lau<sup>1</sup>

Stephen Lin<sup>2</sup>

<sup>1</sup> City University of Hong Kong

nanxuanzhao@gmail.com, rynson.lau@cityu.edu.hk, {wuzhiron,stevelin}@microsoft.com

[http://nxzhao.com/projects/DiLo\\_localization/](http://nxzhao.com/projects/DiLo_localization/)

## Abstract

Recent progress in contrastive learning has revolutionized unsupervised representation learning. Concretely, multiple views (augmentations) from the same image are encouraged to map to the similar embeddings, while views from different images are pulled apart. In this paper, through visualizing and diagnosing classification errors, we observe that current contrastive models are ineffective at localizing the foreground object, limiting their ability to extract discriminative high-level features. This is due to the fact that view generation process considers pixels in an image uniformly. To address this problem, we propose a data-driven approach for learning invariance to backgrounds. It first estimates foreground saliency in images and then creates augmentations by copy-and-pasting the foreground onto a variety of backgrounds. The learning still follows the instance discrimination pretext task, so that the representation is trained to disregard background content and focus on the foreground. We study a variety of saliency estimation methods, and find that most methods lead to improvements for contrastive learning. With this approach (DiLo), significant performance is achieved for self-supervised learning on ImageNet classification, and also for object detection on PASCAL VOC and MSCOCO.

## Introduction

Visual recognition has been revolutionized by deep learning in the fashion of assembling considerable amounts of labeled data (Deng et al. 2009) and training very deep neural networks (Krizhevsky, Sutskever, and Hinton 2012). However, collection of supervisory signals, especially at a very large scale, is constrained by budget and time. Due to this, there has been a growing interest in self-supervised and unsupervised learning which do not face this practical limitation. For high-level visual recognition, previous approaches in self-supervised learning define proxy tasks which do not require human labeling but encode useful priors (Zhang, Isola, and Efros 2016; Doersch, Gupta, and Efros 2015) for object recognition. Recent advances in self-supervised contrastive learning rely on the proxy task of instance discrimination (Dosovitskiy et al. 2015b; Wu et al. 2018), where



Figure 1: Motivation: for natural images with objects, backgrounds are usually shared across categories, while the distinctive region for determining the object is localized.

invariances are encoded and learned from low-level image augmentations such as spatial cropping and color jittering.

In this paper, by visualizing and diagnosing errors made by recent self-supervised contrastive models, we identify a strong pattern which is overlooked by prior works. Specifically, we find that current self-supervised models lack the ability to localize foreground objects, and the learned representation can be predominantly determined by background pixels. This is actually unsurprising, as self-supervised learning generally treats each spatial location as equally important, and it is well known that neural networks are prone to “cheat” (Zhang, Isola, and Efros 2016) by taking advantage of unintended information. As a result, a network cannot be expected to discover objects unless it is driven to do so (Arandjelović and Zisserman 2019).

In supervised visual recognition, localization has been demonstrated to be a strong by-product of training on image-level labels. Strong object localization performance has been shown using the gradient of the class score in the pixel space (Simonyan, Vedaldi, and Zisserman 2013). It has also been found that adding precise localization information does not bring significant gains for PASCAL object classification when transferred from ImageNet (Oquab et al. 2015). Moreover, object segments have been estimated using only image-level labels via a class activation mapping method (Zhou et al. 2016). As suggested in Figure 1, we hypothesize that the learning signal that drives localization comes from the

\*Equal Contributions.

Copyright © 2021, Association for the Advancement of Artificial Intelligence ([www.aaai.org](http://www.aaai.org)). All rights reserved.

category-wise supervisory labels, because background contents (e.g., grass, sky, water) are usually shared among different categories while foreground objects are only salient within the same category.

The gap in the localization ability between self-supervised and supervised models motivates us to explore approaches for distilling localization of self-supervised representations. We study this problem by first estimating a foreground saliency mask for each training image. The training image and its corresponding saliency map are then used to create augmentations by pasting the foreground object onto various backgrounds. During representation learning, we follow recent contrastive representation learning methods using the augmentations for the same object on different backgrounds. This encourages the representation to become invariant to backgrounds, enabling localization of the foreground object.

For generating our augmentations, several saliency estimation methods are examined, including traditional unsupervised techniques (Zhu et al. 2014; Yan et al. 2013; Wei et al. 2012), and a saliency network (Qin et al. 2019). Our model (DiLo) shows consistent improvements of 2% – 6% over the baselines. This clearly demonstrates that object recognition benefits from better localization, and that our approach is effective for solving the localization problem. Due to its better localization ability, we also achieve state-of-the-art transfer learning results for object detection on PASCAL VOC and MSCOCO.

In summary, this paper makes the following contributions:

- 1) A visualization-based study of recent self-supervised contrastive learning models that shows a limited capacity to localize objects.
- 2) A data-driven method that improves the localization ability of contrastive representation learning, demonstrating its effectiveness on both image classification and object detection transfer tasks.
- 3) An investigation of different kinds of saliency estimation methods for improving localization, including traditional saliency and network-predicted saliency.

## Related Work

**Unsupervised and Self-Supervised Learning.** Unsupervised learning aims to extract semantically meaningful representations without human labels (de Sa 1994). Self-supervised learning is a sub-branch of unsupervised learning which automatically generates learning signals from the data itself. These learning signals have been derived from proxy tasks that involve semantic image understanding but do not require semantic labels for training. These tasks have been based on prediction of color (Zhang, Isola, and Efros 2016), context (Doersch, Gupta, and Efros 2015; Pathak et al. 2016), rotation (Gidaris, Singh, and Komodakis 2018), and motion (Pathak et al. 2017). Auto-encoders (Vincent et al. 2008) and GANs (Goodfellow et al. 2014; Donahue and Simonyan 2019) have also shown promising results for representation learning through reconstructing images.

Contrastive learning is another promising direction of work for self-supervised learning. It achieves invariances in a data-driven fashion by image augmentations. Exem-

plar CNN (Dosovitskiy et al. 2015b) and instance discrimination (Wu et al. 2018) create augmentations of an image through changes in color, spatial location and scale. PIRL (Misra and Maaten 2020) and CPC (Oord, Li, and Vinyals 2018) formulate contrastive learning in image patches. CMC (Tian, Krishnan, and Isola 2019) considers explicit modeling of different views. MoCo (He et al. 2019) and SimCLR (Chen et al. 2020a) scale contrastive learning by momentum encoders and large batch sizes. Our paper is in line with these works, and we propose a non-trivial augmentation for distilling localization information.

**Saliency Estimation.** Saliency estimation refers to the task of estimating the locations of interesting objects consistent with human perception. For learning saliency, datasets (Bylinskii et al.) have been collected by tracking eye fixations over an image. Later works usually consider saliency as the full foreground object.

Previous non-learning based approaches (Zhu et al. 2014; Yang et al. 2013) rely on handcrafted features and use priors to find salient object regions. Useful priors include background priors (Han et al. 2014), color contrast priors (Cheng et al. 2014), and objectness (Jiang et al. 2013b). Deep supervised methods (Qin et al. 2019) train a segmentation network to regress the foreground mask, outperforming all non-learning based methods. Recent research on saliency estimation also explores unsupervised learning methods. It integrates multiple non-learning based methods into a noise optimization framework (Zhang, Han, and Zhang 2017), showing results that are on par with supervised methods.

In a network, the salient region corresponds to pixels that fire for the classification decision. Previous works study this in both the input space via gradient visualization (Simonyan, Vedaldi, and Zisserman 2013) and the output space via activation mapping (Zhou et al. 2016). A prior work (Zhou et al. 2014) also finds the salient region by optimizing a minimal region that determines the classification response.

**Copy-and-paste for Visual Recognition.** Several works create data in a copy-and-paste fashion for visual recognition. A key insight of such an approach is that data being generated may not look realistic, but the trained model generalizes surprisingly well to real data. For example, Flying Chairs (Dosovitskiy et al. 2015a) renders chairs onto various backgrounds to generate data for optical flow estimation. Cut-paste-learn (Dwibedi, Misra, and Hebert 2017) randomly puts household object instances in an indoor environment for instance detection and segmentation. Instaboost (Fang et al. 2019) spatially shifts the foreground objects as a means of data augmentation for instance segmentation. Copy-pasting GAN (Arandjelović and Zisserman 2019) uses the copy-and-paste idea to discover objects in an unsupervised manner. However, their experiments are performed on toy examples, such as discovering artificial boxes. Moreover, they do not show how discovering objects may help recognition. Our work follows this path, but in contrast to these previous works our method is targeted to self-supervised representation learning. We note that our augmented images are extremely unrealistic, but provide useful information for learning a recognition model.

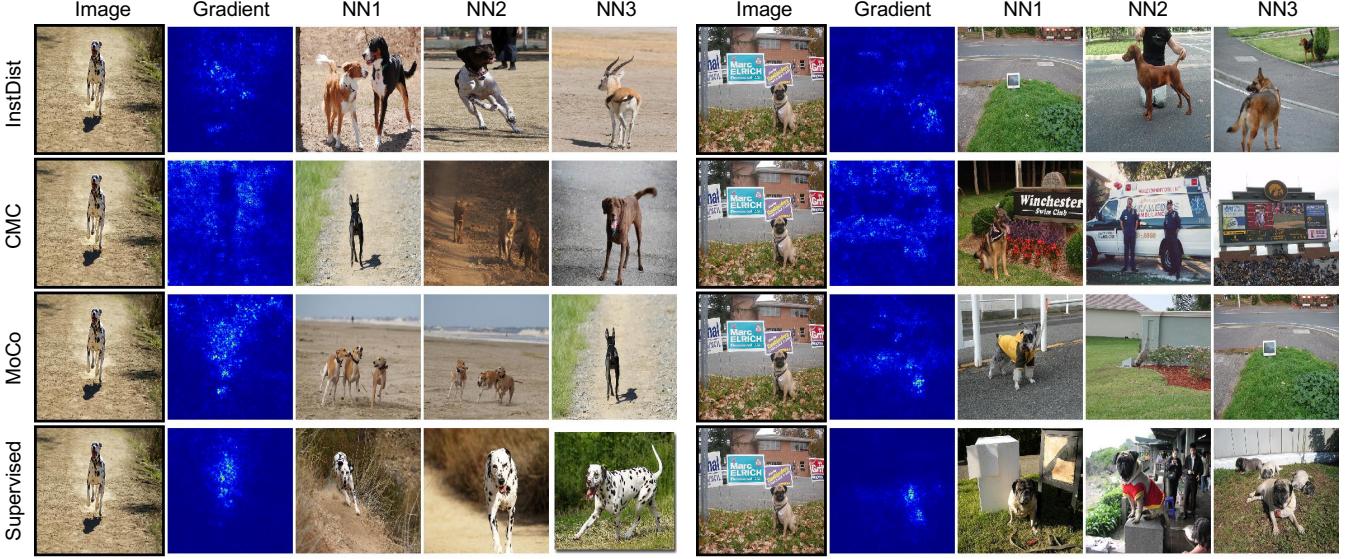


Figure 2: Visualizing and analyzing the error patterns of self-supervised contrastive models. Given an input for each model, we visualize its top-3 nearest neighbors in the embedding space, as well as the gradient on the pixel space with respect to the classification signal. Compared with the supervised model, which is able to localize the salient objects, self-supervised models (InstDisc, CMC, MoCo) look holistically over the image and are prone to distraction by backgrounds. More visualizations are available in the appended supplement.

**Image Augmentations.** Data augmentation plays a key role in visual recognition. Recent works devise handcrafted augmentations (DeVries and Taylor 2017) or learning-based methods (Cubuk et al. 2019; Ratner et al. 2017) to boost representation learning especially in semi-supervised learning. Our copy-paste augmentation is the first introduced for self-supervised learning. From it, we seek to gain further understanding about the ineffective localization problem in self-supervised learning.

## Revisiting Contrastive Learning

Our work builds on recent contrastive learning methods for unsupervised learning, where most work follow the pretext task of instance discrimination. The algorithm first generates image augmentations in the spatial domain, scale space, and color space and then it encourages augmentations of the same image to have similar feature embeddings, and augmentations of different images to have dissimilar embeddings.

Let  $x$  denote the image and  $\mathbf{v} = f(x)$  be the feature embedding, where  $f(\cdot)$  is the embedding function implemented as a convolutional neural network. Let  $\tilde{x} = \mathcal{T}(x)$  represent an augmentation for image  $x$ , where  $\mathcal{T}$  is a random augmentation function. The probability of the augmentation  $\tilde{x}$  to be classified as the  $i$ -th identity is expressed as

$$P(i|\tilde{x}) = \frac{\exp(\mathbf{v}_i^T \tilde{\mathbf{v}}/\tau)}{\sum_{j=1}^n \exp(\mathbf{v}_j^T \tilde{\mathbf{v}}/\tau)}, \quad (1)$$

where  $\tau$  is a temperature parameter and  $n$  is the total number of images in the dataset.  $\tilde{\mathbf{v}} = f(\tilde{x})$ ,  $\mathbf{v}_i = f(x_i)$  are the embeddings for image  $x_i$  and  $\tilde{x}$ . The learning objective is to

minimize the negative log-likelihood over the dataset:

$$J(\theta) = -\sum_{i=1}^n \log P(i|f_\theta(\mathcal{T}(x_i))). \quad (2)$$

Recent self-supervised learning methods such as InstDisc (Wu et al. 2018), CMC (Tian, Krishnan, and Isola 2019), MoCo (He et al. 2019), SimCLR (Chen et al. 2020a) all share a similar formulation. The effectiveness of such an approach for unsupervised learning strongly relies on the types of augmentations  $\mathcal{T}(\cdot)$ , i.e., image transformation priors that do not change object identity. In Table 1, we summarize the role of data-driven augmentations for both a typical self-supervised MoCo ResNet50 model (He et al. 2019) and the supervised model. We gradually add each type of transformation to the set of augmentations. The performance is measured on the ImageNet validation set of 1000 classes, and evaluated by linear classifiers.

We find that the unsupervised representation gains much more classification accuracy from the augmentations than the supervised representation. This indicates that the priors present in the augmentations strongly overlap with the modeling cues from semantic labels. Adding intense color jittering improves the unsupervised representation but hurts the supervised representation. This suggests that the color jitter prior expands beyond the original data distributions. Nevertheless, adding a prior that only partially relates to semantics improves self-supervised learning significantly.

## Visualizing / Diagnosing Contrastive Learning

A variety of methods have been presented for visualizing the behavior of supervised convolutional neural networks, based on deconvolution (Zeiler and Fergus 2014), class-specific

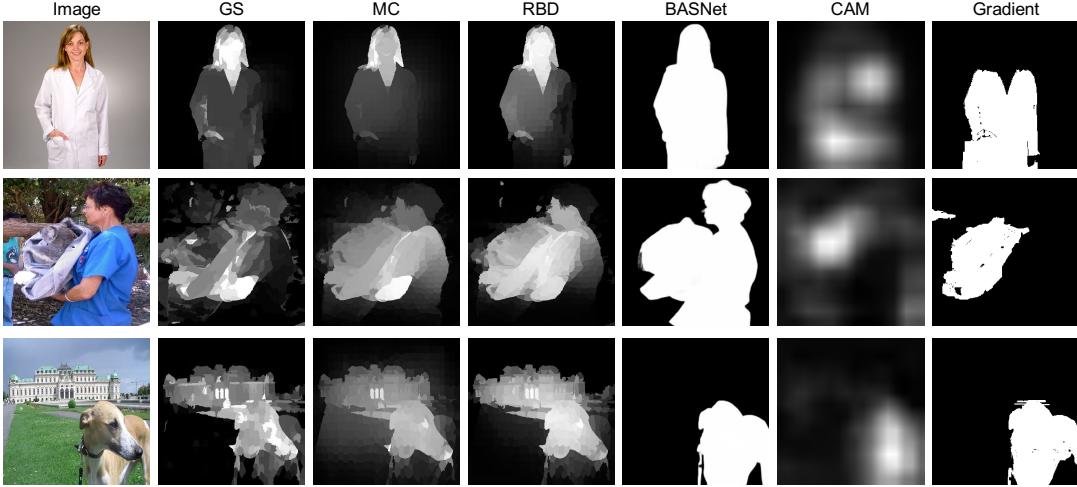


Figure 3: Examples of saliency estimations methods. We show 6 saliency estimations, including traditional methods (GS (Wei et al. 2012), MC (Jiang et al. 2013a), RBD (Zhu et al. 2014)), a network predicted saliency BASNet (Qin et al. 2019), and class-specific methods visualized from a pretrained network (CAM (Zhou et al. 2016), Gradient (Simonyan, Vedaldi, and Zisserman 2013)).

Table 1: A comparison study of the role of data augmentations for learning self-supervised and supervised representations. Please refer to the main text for details.

Augmentations	Self-Supervised	Supervised
+ Flipping	6.4	70.9
+ Spatial Scale Crop	40.4	77.5
+ Color Jitter	56.9	77.4
+ Random Gray	60.6	77.7

gradients (Simonyan, Vedaldi, and Zisserman 2013), and class activation mapping (Zhou et al. 2016; Selvaraju et al. 2017). However, there is little work on visualizing and analyzing the error patterns of self-supervised models, particularly for understanding the relationship between the proxy task and the semantic labels.

In the following, we visualize some representative contrastive learning models with a focus on understanding the salient regions when self-supervised networks make wrong predictions.

**Visualization Methods.** We adapt two visualization methods to our objective.

- Nearest Neighbors. A straightforward way to diagnose what a feature has learned is to find the nearest neighbors in the feature space. By identifying patterns on what draws neighbors close to each other, we gain insights about what the features represent.
- Class-specific gradients. The magnitude of class-score gradients in the pixel space provides information about how important the pixels are for classification. This approach has proven to be strong for weakly-supervised object localization (Simonyan, Vedaldi, and Zisserman 2013). Since self-supervised models do not have classifiers for objects, we train a linear classifier on top of the

extracted features. Then we do back-propagation through the linear classifier and the rest of the self-supervised network to calculate the gradients in the pixel space.

**Investigated Models.** We examine three self-supervised models, including InstDist, CMC, and MoCo.

- InstDist (Wu et al. 2018) treats each individual instance as a class and learns a representation by non-parametric classification with a memory bank implementation.
- CMC (Tian, Krishnan, and Isola 2019) explicitly decouples an image to two views, namely the lightness and color channels. Learning follows to maximize the mutual information between views.
- MoCo (He et al. 2019) follows InstDist and further proposes a momentum encoder to fix the consistency between positives and a queue-based memory for scalability.

**Error Patterns.** Figure 2 illustrates our major findings. We observe that for a considerable number of error cases, the similarity between a query and its nearest neighbors exists mainly in their backgrounds. Gradient-based saliency visualization confirms such findings, as the salient regions for self-supervised models are spread across the background instead of the foreground. For comparison, we also show the corresponding results for the supervised models, which instead show similarities among the foregrounds.

Since these self-supervised methods rely heavily on augmentations to learn invariances, and these augmentations treat foreground and background pixels equally, thus they do not enforce a loss that drives the model to discover objects. This lack of localization ability calls for salient region modeling in self-supervised learning.

## DiLo: Distilling Localization via Background Invariance

Our goal is to learn a representation from which the foreground object can be automatically localized, such that dis-

iscriminative regions can be focused on to improve recognition. We propose to distill the ability for object localization by learning invariance against the background. We first describe methods to extract foreground regions by saliency estimation, and then introduce our background augmentations by copy-and-paste operations.

## Saliency Estimation

In distilling localization ability for self-supervised methods, our approach first estimates saliency masks. A saliency mask should depict the regions most relevant for classifying the object. Typically, it coincides with the foreground object region, as indicated by most saliency datasets (Wang et al. 2017).

Note that recent research on unsupervised saliency estimation has shown promising progress. However, these models (Zhang et al. 2018; Nguyen et al. 2019) heavily rely on ImageNet and semantic segmentation pretraining, which violates our unsupervised experimental protocols. We avoided these methods in this paper, and instead consider the following techniques.

**Traditional Methods.** Traditional saliency estimation methods use handcrafted features, and rely on priors and heuristics to find the dominant object in an image. Useful priors include the background prior (pixels on the image border are more likely to be background) and the color contrast prior (edges with high contrast tend to belong to the foreground). We investigate several high-performing methods: RBD (Zhu et al. 2014), MC (Jiang et al. 2013a), and GS (Wei et al. 2012).

**Saliency Networks.** Recent methods for saliency estimation commonly employ deep learning on annotated saliency datasets (Wang et al. 2017). These deep models outperform traditional methods by a large margin. A state-of-the-art saliency network BASNet (Qin et al. 2019) is included in the investigation, and it is trained on a modest amount of 10K images from scratch.

**Class-specific Saliency.** The aforementioned methods estimate saliency as foreground object regions. However, it is not clear that this represents the discriminative part of an image (e.g., only the face of a person may be important for recognizing humans). To keep the problem open, we also compare with CAM (Zhou et al. 2016) and a gradient-based method (Simonyan, Vedaldi, and Zisserman 2013) through class-specific visualizations. For (Simonyan, Vedaldi, and Zisserman 2013), we convert the gradients to a mask using a segmentation algorithm (Gulshan et al. 2010).

**Summary.** Figure 3 shows examples of the saliency visualizations. Traditional methods are seen to be noisy, while network-produced saliency is much cleaner. It can be noticed that class-specific saliency from a pretrained network tends to be more compact around discriminative regions. This indicates that the use of full foreground saliency may not be ideal.

## Copy-and-paste for Background Augmentation

Based on the previous findings, we propose to copy the foreground object estimated from the saliency methods in prior

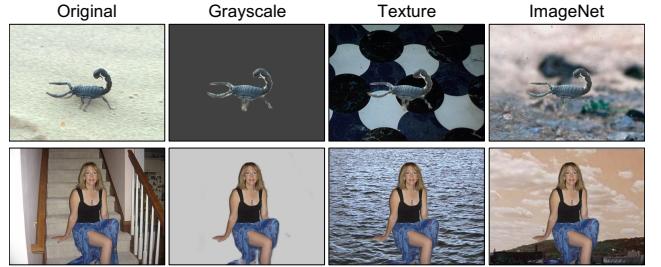


Figure 4: Generated copy-paste augmentations using three kinds of background images.

section , and paste that onto various backgrounds as a means of data-driven augmentation for learning localization.

**Background Datasets.** For this augmentation, we ablate three types of backgrounds.

- Homogeneous grayscale images with a random grayscale level.
- Texture images from the MIT Vision Texture dataset (MediaLab 1995).
- Image crops from ImageNet which have no saliency response using RBD (Zhu et al. 2014).

Figure 4 shows copy-and-pasted examples using various background images.

**Blending.** For pasting, we examine three techniques: directly copying the foreground object onto the background, copying with Gaussian blending on the object borders, and a mixture of the two approaches.

**Accounting for Context.** Context plays an important role in recognizing objects (Torralba 2003). Though the surrounding context of an object may not be the most discriminative region for recognition, it may help to prune the set of candidates. For example, a tree is unlikely to be completely encompassed by sky. To account for this during augmentation, we set a probability of keeping the original full image without copy-and-paste augmentation.

**Integrating other Augmentations.** Since copy-paste augmentation is orthogonal to other previous augmentations, i.e. random scaling, cropping, color jittering, the order of copy-paste augmentation with respect to other augmentations does not matter. In our implementation, we first run copy-paste augmentation to replace the background, and then perform other augmentations.

## Experiments

We conduct a series of experiments on model designs for self-supervised representation learning and their transfer learning abilities.

### Ablation Study

In this section, we first validate our data-driven approach of distilling localization through a series of ablation experiments for image classification on ImageNet.

**Baseline Settings.** Due to its state-of-the-art performance, we largely follow MoCo (He et al. 2019) settings as our

Table 2: Ablation studies for investigating copy-and-pasting augmentations: (a) on various saliency estimation methods (b) on controlling the ratio of using copy-and-pasting augmentation (c) on various background images (d) on blending options.

Saliency	(a)				(b)			(c)			(d)		
	$F_\beta$	MAE	Acc	$\Delta$	Aug Ratio	Linear	$\Delta$	Background	Linear	$\Delta$	Blending	Linear	$\Delta$
MoCo	-	-	60.6	-	MoCo	60.6	-	MoCo	60.6	-	MoCo	60.6	-
GS	0.557	0.173	62.7	+2.1	30%	62.8	+2.2	Texture	60.6	+0.0	No blend	62.4	+1.8
MC	0.627	0.186	62.1	+1.5	50%	62.2	+1.6	Imagenet	62.1	+1.5	Gaussian	62.5	+1.9
RBD	0.630	0.144	62.8	+2.2	70%	61.6	+1.0	Grayscale	62.8	+2.2	Mix	62.8	+2.2
BASNet	0.805	0.056	65.0	+4.4	100%	47.6	-13.0						

baseline for ablation. Specifically, we use a temperature  $\tau = 0.07$  in Eqn. 1, and an embedding dimension of  $D = 128$  for each image. A memory queue (He et al. 2019) of size  $k = 65536$  negatives is used to accelerate discrimination. Training takes 200 epochs with an initial learning rate of 0.03 that is decayed 1/10 at epochs 120 and 160. All models are trained using the ResNet50 architecture and reported on the ImageNet validation set. Performance is evaluated by the linear readoff on the penultimate layer features. The optimization takes 100 epochs and starts with a learning rate of 30 that is decayed every 30 epochs.

**A naive approach.** First of all, to demonstrate the necessity of a data-driven approach, we consider a naive approach that pools the final layer features by masking according to saliency. With this, the performance decreases sharply by 19%, possibly because the model loses too much context. Moreover, by masking out the features, the model is still unable to localize the discriminative regions automatically.

**Saliency Estimation.** In Table 2 (a), we examine several class-agnostic saliency estimation methods. All of them are found to improve performance, even the noisy traditional approaches RBD (Zhu et al. 2014), MC (Jiang et al. 2013a) and GS (Wei et al. 2012). RBD improves the performance by 2.2% and the saliency network BASNet by 4.2%. The supervised BASNet (Qin et al. 2019) is trained on the DUTS dataset (Wang et al. 2017) from scratch with 10,053 training images, which is less than 1% of ImageNet. This indicates potential room for developing better unsupervised saliency approaches. In Table 2, we find a correlation between the saliency performance on the saliency benchmark (by  $F_\beta$  and MAE on DUT-OMRON dataset (Yang et al. 2013)) and the self-supervised representation learning. Better saliency translates to better representations.

**Background Images.** We ablate the use of various background images in Table 2 (c). Texture backgrounds improve the performance very marginally. This is possibly because textured images in the dataset (Bylinskii et al.) are outside of the ImageNet distribution. Homogeneous grayscale backgrounds and ImageNet backgrounds perform similarly well.

**Amount of Augmentation.** During dataloading, we only randomly add copy-and-paste augmentations with a probability ratio. We ablate the ratio in Table 2 (b). With only 30% to 50% of images receiving copy-and-pastes, we significantly improve the performance by 2% – 4%. Always using the copy-and-paste augmentation hurts performance.

**Blending Options.** When copy-and-pasting an object to a background, blending has proven to be important for object detection (Dwibedi, Misra, and Hebert 2017). In our study

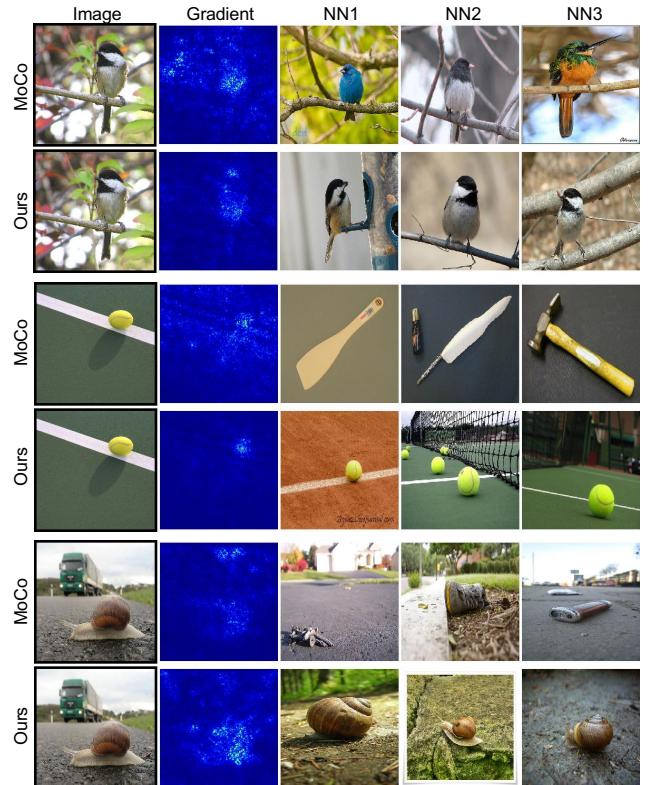


Figure 5: Successful examples where our model outperforms our baseline. The improvement is due to better localization and background invariance.

in Table 2 (d), blending appears to improve the performance minorly about 0.4%. This difference is possibly because detection requires realistic boundaries, which prevents the network from taking shortcuts, while for classification, boundary cheating is not as significant.

**Visualizations.** In Figure 5 and Figure 6, we visualize examples where our model outperforms the baseline, as well as some failure cases. For all the successful cases, our salient region on the gradient and the nearest neighbors focus on the discriminative object, while the baseline approach is distracted by the background. This validates the claim that our data-driven augmentation drives the model to learn to automatically localize the object. Such localization leads to better recognition performance.

For the failure cases, we compare our model with the supervised model. We find that there are two error patterns.

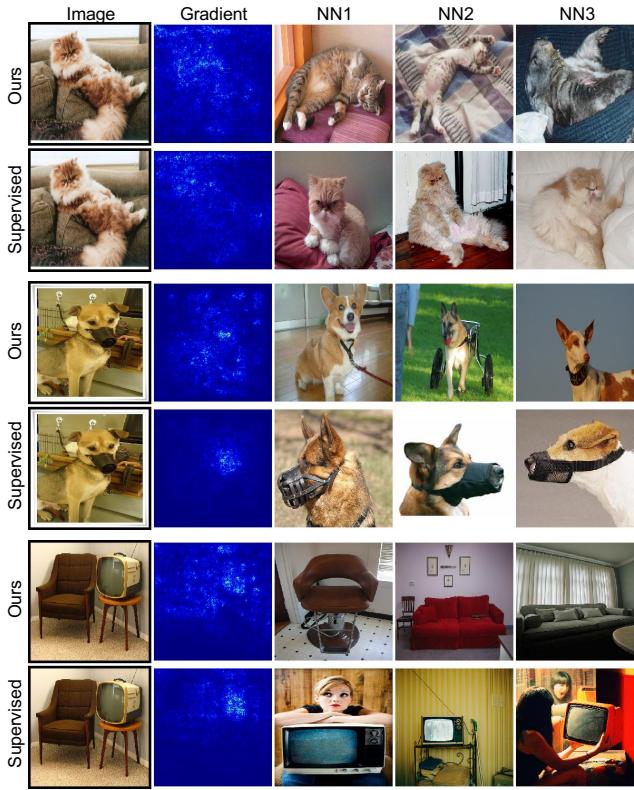


Figure 6: Failures where our model underperforms the supervised model. The model finds it difficult when multiple objects appear in the image, or the object is of a fine-grained category.

Table 3: Distilling localization on various contrastive representation learning models for ImageNet classification.

Methods	Original	DiLo-RBD	DiLo-BasNet
InstDist	56.5	59.3	62.9
CMC	63.4	65.0	66.9
MoCo-v1	60.6	62.8	65.0
MoCo-v2	67.5	67.9	69.2

First, multiple objects appear in a single image, and our model makes wrong decisions on where to focus. Second, the testing image is of a fine-grained class, too difficult to recognize without labels.

## Transfer Learning Results

We evaluate the transfer learning ability of our model on object recognition, and object detection benchmarks, and compare with the state-of-the-art methods.

**ImageNet Classification.** We conduct a plug-and-play of DiLo into existing contrastive learning frameworks. Methods being investigated include InstDist (Wu et al. 2018), CMC (Tian, Krishnan, and Isola 2019), MoCo (He et al. 2019) and MoCo-v2 (Chen et al. 2020b). In Table 3, DiLo consistently improves image classification on all baselines. Foreground masks estimated from BasNet are more beneficial than RBD. The results demonstrate that DiLo is orthog-

Table 4: Transfer learning for object detection on VOC 0712. We present the gap to ImageNet supervised pre-training in the brackets for reference. All numbers are the averages of three independent runs.

Method	$AP$	$AP_{50}$	$AP_{75}$
Supervised	53.5	-	81.3
MoCo	55.9 (+2.4)	81.5 (+0.2)	62.6 (+3.8)
DiLo-RBD	56.5 (+3.0)	81.9 (+0.6)	63.3 (+4.5)
DiLo-BasNet	56.9 (+3.4)	82.1 (+0.8)	64.1 (+5.3)

Table 5: Transfer learning for object detection and instance segmentation on COCO. Model is finetuned with Mask-RCNN ResNet50-FPN pipeline and 1x schedule.

Method	$AP^{bb}$	$AP_{50}^{bb}$	$AP_{75}^{bb}$	$AP^{mk}$	$AP_{50}^{mk}$	$AP_{75}^{mk}$
Supervised	39.7	59.5	43.3	35.9	56.6	38.6
MoCo	39.4	59.1	42.9	35.6	56.2	38.0
DiLo-RBD	39.8	59.5	43.3	36.0	56.7	38.6
DiLo-BasNet	40.1	60.0	44.0	36.3	56.8	39.0

onal to prior contrastive learning works.

**Object Detection on PASCAL VOC.** We transfer our pretrained model to object detection by finetuning it on PASCAL VOC 2007+2012 trainval and evaluating on the VOC 2007 test set. Following the state-of-the-art method MoCo (He et al. 2019), we use the exact same training protocol to finetune the Faster R-CNN with a Res50-C4 backbone as with the supervised counterpart. A critical BN layer is added after the conv5 stage in the box prediction head. During training, we finetune all layers with synchronized BN. The finetuning takes 9k iterations. Results are summarized in Table 5. DiLo-RBD and DiLo-BasNet consistently outperform supervised baseline and MoCo on all metrics, especially on  $AP_{75}$  which heavily reflects localization ability.

**Object Detection and Instance Segmentation on COCO.** We transfer the pretrained DiLo model for object detection and instance segmentation on MSCOCO by finetuning it with the Mask-RCNN Res50-FPN pipeline using the Detectron2 codebase. Finetuning takes the default 1x schedule. DiLo-RBD and DiLo-BasNet consistently outperform the MoCo baseline with good margins on both detection and segmentation.

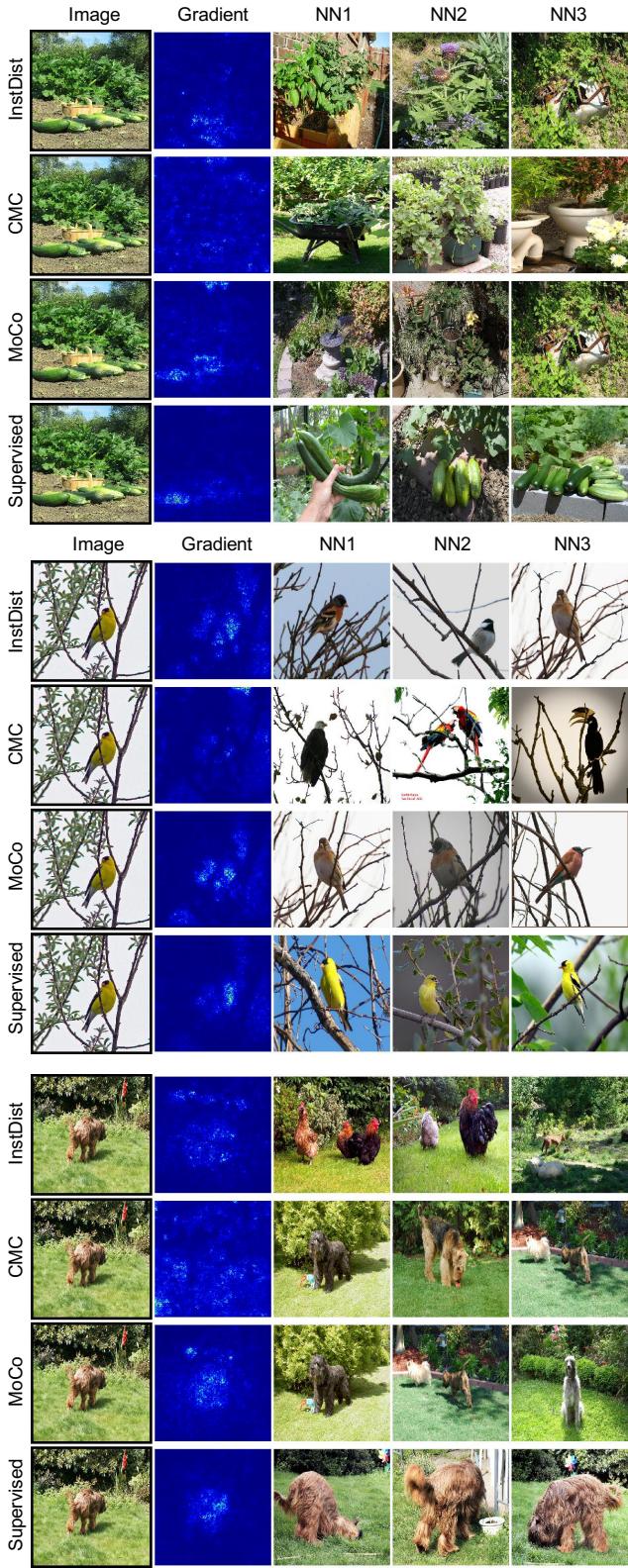
## Conclusion

In this work, we identified a strong error pattern among self-supervised models in their failure to localize foreground objects. We then propose a simple data-driven approach to distill localization via learning invariance against backgrounds. We achieve strong results on ImageNet classification and its transfer performance for object detection on PASCAL VOC 2007. The improvements achieved suggest that the localization problem for self-supervised representation learning is prevalent. However, our method may not be the ideal way to solve this localization problem. We are interested in finding a clever “proxy task” which can help distill such localization abilities.

## References

- Arandjelović, R.; and Zisserman, A. 2019. Object Discovery with a Copy-Pasting GAN. *arXiv preprint arXiv:1905.11369*.
- Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; and Torralba, A. ???? MIT Saliency Benchmark.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.
- Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H.; and Hu, S.-M. 2014. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 113–123.
- de Sa, V. R. 1994. Learning classification with unlabeled data. In *Advances in neural information processing systems*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee.
- DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Donahue, J.; and Simonyan, K. 2019. Large scale adversarial representation learning. *arXiv preprint arXiv:1907.02544*.
- Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015a. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*.
- Dosovitskiy, A.; Fischer, P.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2015b. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*.
- Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Fang, H.-S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.-L.; and Lu, C. 2019. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. *arXiv preprint arXiv:1908.07801*.
- Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*.
- Gulshan, V.; Rother, C.; Criminisi, A.; Blake, A.; and Zisserman, A. 2010. Geodesic star convexity for interactive image segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.
- Han, J.; Zhang, D.; Hu, X.; Guo, L.; Ren, J.; and Wu, F. 2014. Background prior-based salient object detection via deep reconstruction residual. *IEEE Transactions on Circuits and Systems for Video Technology*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.
- Jiang, B.; Zhang, L.; Lu, H.; Yang, C.; and Yang, M.-H. 2013a. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*.
- Jiang, P.; Ling, H.; Yu, J.; and Peng, J. 2013b. Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proceedings of the IEEE international conference on computer vision*.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.
- MediaLab, M. 1995. VisTex texture database. Web site: <http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html>.
- Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6707–6717.
- Nguyen, T.; Dax, M.; Mummadi, C. K.; Ngo, N.; Nguyen, T. H. P.; Lou, Z.; and Brox, T. 2019. DeepUSPS: Deep Robust Unsupervised Saliency Prediction via Self-supervision. In *Advances in Neural Information Processing Systems*.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Pathak, D.; Girshick, R.; Dollár, P.; Darrell, T.; and Hariharan, B. 2017. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2701–2710.
- Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by

- inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. BASNet: Boundary-Aware Salient Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Ratner, A. J.; Ehrenberg, H.; Hussain, Z.; Dunnmon, J.; and Ré, C. 2017. Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, 3236–3246.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive Multiview Coding. *arXiv preprint arXiv:1906.05849*.
- Torralba, A. 2003. Contextual priming for object detection. *International journal of computer vision*.
- Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM.
- Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *CVPR*, 136–145.
- Wei, Y.; Wen, F.; Zhu, W.; and Sun, J. 2012. Geodesic saliency using background priors. In *European conference on computer vision*. Springer.
- Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer.
- Zhang, D.; Han, J.; and Zhang, Y. 2017. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Zhang, J.; Zhang, T.; Dai, Y.; Harandi, M.; and Hartley, R. 2018. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*. Springer.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.



## Visualizing and Diagnosing Self-Supervision

We provide more visualization examples in the following figures. Given an input for each model, we visualize its top-3 nearest neighbors in the embedding space, as well as the gradient on the pixel space with respect to the classification signal. Compared with the supervised model, which is able to localize the target objects, self-supervised contrastive models (InstDisc, CMC, MoCo) look holistically over the image and are prone to distraction by backgrounds.

