Concurrent Discrimination and Alignment for Self-Supervised Feature Learning

Anjan Dutta[1], Massimiliano Mancini[2], Zeynep Akata[2]
[1]University of Exeter, [2]University of Tübingen

Abstract

Existing self-supervised learning methods learn representation by means of pretext tasks which are either (1) discriminating that explicitly specify which features should be separated or (2) aligning that precisely indicate which features should be closed together, but ignore the fact how to jointly and principally define which features to be repelled and which ones to be attracted. In this work, we combine the positive aspects of the discriminating and aligning methods, and design a hybrid method that addresses the above issue. Our method explicitly specifies the repulsion and attraction mechanism respectively by discriminative predictive task and concurrently maximizing mutual information between paired views sharing redundant information. We qualitatively and quantitatively show that our proposed model learns better features that are more effective for the diverse downstream tasks ranging from classification to semantic segmentation. Our experiments on nine established benchmarks show that the proposed model consistently outperforms the existing state-of-the-art results of self-supervised and transfer learning protocol. Code can be found at `https://github.com/AnjanDutta/codial`.

1. Introduction

Supervised deep learning methods achieve a high performance only when trained on a large amount of labeled data gathered through expensive and error-prone annotation procedure. Therefore, depending on the training procedure and underlying dataset, supervised learning might yield features that mainly focus on the local statistics, and hence might end up learning spurious correlations. Since it is well known that global statistics possess better generalization capability [17], supervised learning often suffers from a poor generalization ability. To alleviate these limitations [19], and to learn representations from large-scale unlabeled datasets, self-supervised learning (SSL) [4, 5, 30] has been proposed. One way to do that is to learn some discriminative tasks, such as the recognition of rotation angle [10] or a local transformation [16, 17] that needs understanding global statistics. Another strategy is to perform diverse image transformations for a single image, and then aligning them with some criterion [15, 25, 2, 12, 7, 34]. In other words, the first group of methods specifies which features should be repelled from each other but does not describe explicitly which features should be attracted to each other (see Figure 1, upper left). On the contrary, the second group specifies which features should be attracted, but does not set out which features should be repelled (see Figure 1, upper right), which creates ambiguities in feature learning. Although [8] proposes to recognize angle of image rotations and simultaneously minimizing the Euclidean distance between representations of the transformed images and their mean, this solution could yield trivial solution by simply learning the same zero vector as the image feature [18].

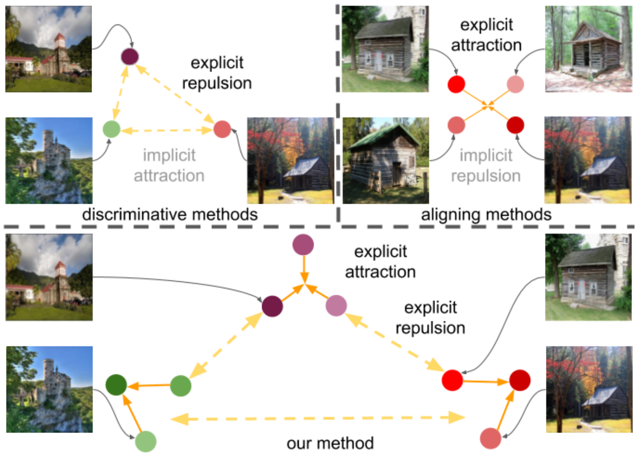

Figure 1. Self-supervised learning methods are (1) discriminative, *i.e*. explicitly specify which features should be repelled, and (2) aligning, *i.e*. indicate which features should be attracted. Our method unifies the advantages of both the categories and explicitly formulates the features to be attracted and repelled simultaneously.

To address this fundamental shortcoming of SSL methods, we propose CODIAL (COncurrent DIscrimintion and ALignment), a novel method that combines the benefit of discriminative and aligning methods. The pre-training task of our method is to discriminate the image transformations and concurrently align global image statistics. For doing that, we define primary and auxiliary image transformations which are orthogonal to each other and respectively use them to discriminate and align global image statistics. Motivated by [17], we choose the primary transformations in such a way that the local image statistics remain largely unchanged, such as recognition of image rotation angles, warping transformations. The auxiliary transformations are the standard ones for image augmentations, such as random crop, horizontal flip, color jittering, blurring etc [2, 12]. The repulsion is achieved via a discriminative classifier and the feature alignment is done by engaging a Jensen-Shannon mutual information estimator [14, 7].

In this work, we make the following contributions: (1) We introduce a novel multi-task method for self-supervised feature learning which combines the advantages of discriminative and alignment based works; (2) Our method successfully avoids degenerating and shortcut solutions by its design which we further enforce by introducing information bottleneck regularizer term to our learning objective; (3) We perform extensive experiments on nine benchmark datasets and achieve state-of-the-art performance both in self-supervised and transfer learning experiments.

2. Related Work

In this section, we review the relevant literature of discriminative and alignment-based semi-supervised learning as well as hybrid methods.

Discriminative Methods: Majority of discriminative methods rely on using auxiliary handcrafted prediction tasks to learn their representation. Early works within this group try to determine spatial configuration of patches [4, 27, 29, 26], which needs distinguishing local patches. Since local features are not sufficient to learn spatial configuration, these pretext tasks necessarily learn global statistics up to the size of local patch or tile. Balancing the local and global features has also been seen in obtaining and matching the local and global statistics [28]. The pretext task of predicting rotation [10] distinguishes the features of the images rotated by 0°, 90°, 180°, 270°, which essentially learns a representation suitable to recognize the angle. The pretext task involved in [16] uses adversarial training to distinguish real images from images with synthetic artifacts for learning visual representation. Few other works use pseudo labels obtained from an intermediate unsupervised clustering step for the networks [1, 9]. Recently, Jenni *et al.* [17] extend the work of Gidaris *et al.* [10] and propose to discriminate local transformation together with some previously proposed predictive tasks. In this work, we use the rotation and warping as the primary transformations for using in the predictive task. Different to the prior works, we randomly crop the original image maintaining the specification mentioned in Section 4 before generating each transformation, *i.e.* we apply the rotation and warping transformations on the cropped version of the original image, which has been proven to be effective for our model.

Alignment Methods: Early works within this group apply some image transformations and align the transformed and real image for learning the underlying model. Inpainting [31] and colorization [36, 37] are among the first few approaches that belong to this category. In case of inpainting, the pretext task involves removing a set of pixels as transformation and then reconstructing the missing part. The pretext task required in colorization [36] removes color as transformation and recovers the color information. Since both the methods map the images invariant to the transformation, they are considered as the alignment methods. Recently proposed vast majority of methods within this category are based on contrastive learning [30, 13, 15, 14, 2, 7, 25, 34, 12], which generally avoids defining pretext tasks, and instead focuses on bringing representation of different views obtained by applying different transformations to the same image closer ('positive pairs'), and implicitly separating representations from different images ('negative pairs'). Contrastive methods often require comparing each example with many other examples to work well [13], which essentially indicates the importance of negative sampling. Contrastively, in this paper, we align different views or transformations by maximizing mutual information [14, 7] which is achieved by maximizing the lower bound of Jensen-Shenon divergence [14]. Additionally, to avoid shortcut or degenerating solution, we impose a bottlenecking regularization term inspired by information theory [35, 7], which boosts our performance.

Hybrid Methods: Few other works have explored the combination of discriminative and alignment methods. Among them, Feng *et al.* [8] show that a combination of the rotation prediction task [10] and the alignment task through minimizing the representation distance with squared loss achieves encouraging results. Different to the existing hybrid methods, in this paper, we maximize mutual information as a way to align transformed images. Additionally, we use cropped transformations for the predictive tasks, which has also been proven to be competent in our experiment.

Our work is a hybrid method, where we define variant of transformations including crop, rotation, warp, blur, and discriminate only a subset of them. We ensure the transformations to contain mutually redundant information, which allow maximizing the mutual information. Addition to the discriminative task, we maximize their mutual information which is not same as merely minimizing representation distance; the presence of entropy within mutual information avoids degeneracy, as discussed in the following.

3. Concurrent Discrimination and Alignment

Our CODIAL framework learns unsupervised image representations by jointly solving two different pretext tasks: (1) recognizing the variations in global image statistics underwent by primary image transformations and (2)

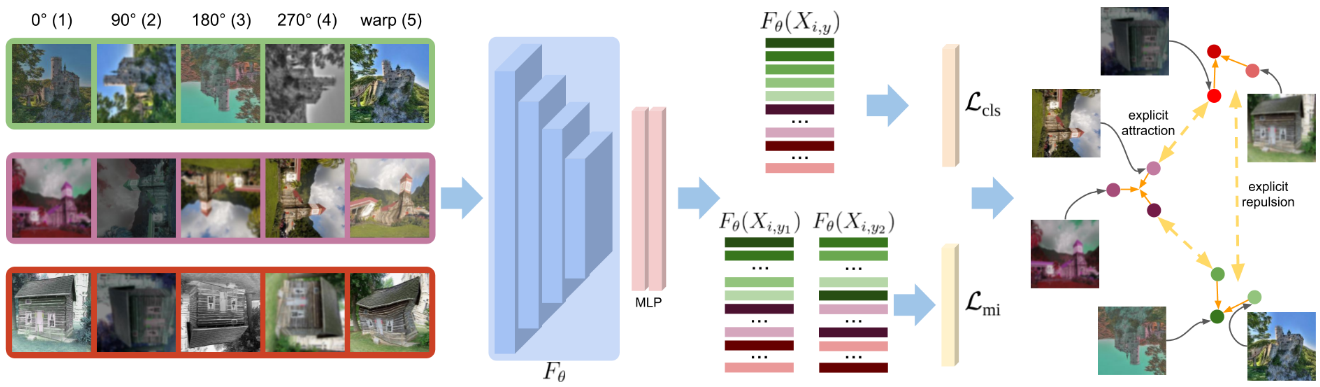

Figure 2. In our CODIAL model, we learn image representations F_θ. In addition to auxiliary image transformations, we learn representations $F_\theta(X_{i,y})$ from primary transformations $X_{i,y}$ of images X_i. $F_\theta(X_{i,y})$ are directly utilized to train the model using the classification objective $\mathcal{L}_{\text{cls}}$ and they are paired to form tuples $(F_\theta(X_{i,y_1}), F_\theta(X_{i,y_2}))$ used for maximizing the mutual information criteria $\mathcal{L}_{\text{mi}}$. The joint discriminating and aligning criteria explicitly specifies which features should be repelled and which ones should be attracted.

maximizing the mutual information of the image pairs formed with the transformed ones. Figure 2 depicts a comprehensive pipeline of our model considering images underwent with primary and auxiliary transformations $X_{i,y}$.

Our network model F_θ results in the transformed representations $F_\theta(X_{i,y})$ for each of the transformed images $X_{i,y}$ which are directly classified and trained with the objective in eqn. (1). The transformed representations $F_\theta(X_{i,y})$ are paired to form tuples $(F_\theta(X_{i,y_1}), F_\theta(X_{i,y_2}))$ to maximize the mutual information by minimizing the objective in eqn. (6). The joint discrimination and alignment strategy only focuses on the generic object features, such as gradient, color, texture etc. which can be distinguished through transformations and are also common across views.

3.1. Predicting Image Transformations

In addition to the primary transformations in [17], *i.e.* rotation and warping, we use horizontal flipping, blurring, cropping and color jittering as auxiliary image transformations for creating deformed images so that the downstream representation becomes more robust to those distortions.

Given a dataset of unlabeled images $S = \{X_i\}_{i=1}^N$ and a set of primary image transformations $T = \{t(X, y)\}_{y=1}^K$ for each image, we define i-th image with the y-th transformation as $X_{i,y} = t(X_i, y)$. A neural network model F_θ is trained to classify each transformed image to one of the K primary transformations classes optimizing the following objective:

$$\mathcal{L}_{\text{cls}} = \min_\theta \frac{1}{KN} \sum_{i=1}^{N} \sum_{y=1}^{K} \ell_{\text{cls}}(F_\theta(X_{i,y}); y) \tag{1}$$

where ℓ_{cls} is the standard parametric cross-entropy loss for a multi-class classification problem. Particularly, we consider the images rotated with angles $0°$, $90°$, $180°$, $270°$ respectively as classes 1, 2, 3 and 4, and the warped image as class 5. Note that the auxiliary transformations are applied before the primary transformations. Particularly, random horizontal flipping is applied to the original image and the rest of the auxiliary transformations are applied before creating the individual transformed classes. Therefore note that the transformed images are not exactly the transformed version of the same image content. Intuitively, these variations in image transformations make the recognition task more difficult, which effectively makes the downstream representation more robust.

Predicting rotation is particularly effective for most natural images having objects in an up-front posture because any rotation of the image will result in an unusual object orientation. However, despite its simplicity and effectiveness, the assumption of rotation prediction task fails for the rotation-agnostic images [8]. Similarly, an image warping is a smooth deformation of the image coordinates defined by n pixel coordinates $\{(a_k, b_k)\}_{k=1,\ldots,n}$, which act as control points as represented in the following transformation

$$\begin{bmatrix} a_k + \Delta a_k \\ b_k + \Delta b_k \end{bmatrix} = \mathbf{T} \begin{bmatrix} a_k \\ b_k \end{bmatrix} \tag{2}$$

where Δa_k and Δb_k are the offsets and are uniformly sampled from a range $[-d, d]$, where $d = 0.1 \times \text{dimension}(X_i)$. Image warping effectively changes the local image statistics only minimally and makes it difficult to distinguish a warped patch from a patch undergoing a change in perspective. Hence, the classifier needs to learn global image statistics to detect image warping.

The main motivation behind the predictive classification task is to learn the generic features that distinguish those transformations and thus, effectively perform various downstream tasks. Image transformations, such as rotation [10, 8], warping [16] change the global image statistics while the local visual information is kept intact. However, a network trained to solve a self-supervised task might accomplish it by using local statistics [4, 17]. Such solutions are usually called *shortcuts* and are a form of degenerate learning as they learn features with poor generalizations capabilities. We avoid those degenerate solutions by maximizing the mutual information between all possible pairs of transformed images as described next.

3.2. Maximizing Mutual Information

Mutual information (MI) measuring relationship between random variables has often been used to learn the representation F_θ. Let X_{i,y_1}, X_{i,y_2} be a paired data sample from a joint probability distribution $P(X_{i,y_1}, X_{i,y_2})$, which is ensured as X_{i,y_1} and X_{i,y_2} are the two different transformed versions of the same image X_i. A representation F_θ can be learned by maximizing the mutual information between the encoded variables:

$$\max_\theta I(F_\theta(X_{i,y_1}); F_\theta(X_{i,y_2})) \tag{3}$$

where I denotes the mutual information and eqn. (3) is equivalent to maximizing the predictability of $F_\theta(X_{i,y_1})$ from $F_\theta(X_{i,y_2})$, and vice versa. The aim of eqn. (3) is to align the representations of paired samples. However, it is not the same as minimizing representation distance. In other words, the presence of entropy in the case of mutual information I allows to avoid degenerate solutions.

We use the Jensen-Shannon mutual information estimator [14, 7], which is a sample based differentiable mutual information lower bound. This procedure of maximizing MI needs introducing an additional parametric critic $C_\xi(F_\theta(X_{i,y_1}), F_\theta(X_{i,y_2}))$ which is jointly optimized with other parameters during the training procedure using re-parameterized samples from $F_\theta(X_{i,y_1})$ and $F_\theta(X_{i,y_2})$. Now, eqn. (3) can trivially be solved by setting F_θ to the identity function because of the data processing inequality:

$$I(X_{i,y_1}; X_{i,y_2}) \geq I(F_\theta(X_{i,y_1}); F_\theta(X_{i,y_2})). \tag{4}$$

To alleviate the trivial solution, we impose a KL-divergence based regularizer rooted from the information bottleneck principle [35] which theoretically aligns the representations from $F_\theta(X_{i,y_1})$ to $F_\theta(X_{i,y_2})$, and vice versa, and discards transformation specific information as much as possible:

$$\begin{aligned} R_{\text{mib}}(F_\theta(X_{i,y_1}), F_\theta(X_{i,y_2})) = \frac{1}{2} D_{\text{KL}}(F_\theta(X_{i,y_1}) || F_\theta(X_{i,y_2})) \\ + \frac{1}{2} D_{\text{KL}}(F_\theta(X_{i,y_2}) || F_\theta(X_{i,y_1})) \end{aligned} \tag{5}$$

where D_{KL} denotes the KL divergence for joint observation between two views. The above regularizer term when combined with the mutual information maximization objective, results in the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{mi}} = & - I(F_\theta(X_{i,y_1}); F_\theta(X_{i,y_2})) \\ & + \beta R_{\text{mib}}(F_\theta(X_{i,y_1}), F_\theta(X_{i,y_2})) \end{aligned} \tag{6}$$

where β is the weight on the regularizer R_{mib} and is increased during training from the initial value 10^{-6} to the final value 1.0 with an exponential scheduling. Hence the final loss function that we aim to minimize is:

$$\mathcal{L} = \min_\theta \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{mi}} \mathcal{L}_{\text{mi}} \tag{7}$$

where λ_{cls} and λ_{mi} are respectively the weights on the classification and mutual information estimation criterion. For creating the pairs essential for estimating and maximizing mutual information, we consider all possible paired combinations of K different transformations of the same image, which results in total ${}^KC_2$ pairs. For efficiency, we uniformly sample a subset of size k ($\leq {}^KC_2$) of such paired combinations. On the other hand, we empirically show that considering large number of such combinations increases the performance of our model. The critic network that estimates mutual information is adversarial and requires negative data points uniformly sampled from the transformed images different from the positive ${}^KC_2$ pairs.

In addition to maximizing MI, we use information bottleneck regularizer eqn. (5) [7] which discards superfluous information between the paired views. We carefully select auxiliary transformations, such as crop, blur, color jitter etc. for creating multiple views without affecting the label information. The transformed positive pairs are supposed to be mutually redundant. Intuitively, a view X_{i,y_1} is redundant with respect to a second view X_{i,y_2} whenever it is irrelevant for the label if X_{i,y_2} is already observed, *i.e.* in terms of mutual information $I(F_\theta(X_{i,y_1}); X_{i,y_1} | X_{i,y_2}) = 0$. With the chain rule of MI:

$$\begin{aligned} I(X_{i,y_1}; F_\theta(X_{i,y_1})) = I(X_{i,y_1}; F_\theta(X_{i,y_1}) | X_{i,y_2}) \\ + I(X_{i,y_2}; F_\theta(X_{i,y_1})) \end{aligned} \tag{8}$$

where it is clear that $I(X_{i,y_1}; F_\theta(X_{i,y_1}))$ can be reduced by minimizing $I(X_{i,y_1}; F_\theta(X_{i,y_1}) | X_{i,y_2})$ as the information $F_\theta(X_{i,y_1})$ contains is unique to X_{i,y_1} and is not predictable by observing X_{i,y_2}. However, since we assume that mutual redundancy between X_{i,y_1} and X_{i,y_2}, $I(X_{i,y_1}; F_\theta(X_{i,y_1}) | X_{i,y_2})$ is irrelevant and can be safely discarded, our first loss:

$$\begin{aligned} & \mathcal{L}_1(\theta, \lambda_1) = \\ & I(F_\theta(X_{i,y_1}); X_{i,y_1} | X_{i,y_2}) - \lambda_1 I(X_{i,y_2}; F_\theta(X_{i,y_1})) \end{aligned} \tag{9}$$

where λ_1 is the Lagrangian multiplier introduced by the constrained optimization. Our second loss:

$$\begin{aligned} & \mathcal{L}_2(\theta, \lambda_2) = \\ & I(F_\theta(X_{i,y_2}); X_{i,y_2} | X_{i,y_1}) - \lambda_2 I(X_{i,y_1}; F_\theta(X_{i,y_2})) \end{aligned} \tag{10}$$

By re-parametrizing the Lagrangian multipliers, the average of the two loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ can be proven to be upper bounded as in eqn. (6) [7]. This allows learning invariances directly from the augmented data, rather than requiring them to be built into the model architecture.

	STL-10					CelebA					CIFAR-10	CIFAR-100	Tiny-ImageNet
Method	c1	c2	c3	c4	c5	c1	c2	c3	c4	c5			
RotNet [10]	58.2	67.3	69.3	69.9	70.1	70.3	70.9	67.8	65.6	62.1	62.1	33.2	23.7
Decouple [8]	59.0	68.6	70.8	72.5	74.6	71.4	73.8	72.7	72.4	72.2	65.9	37.7	28.5
GlobStat [17]	59.2	69.7	71.9	73.1	73.7	71.8	74.0	73.5	72.5	69.2	68.1	39.2	31.2
CODIAL (ours)	**60.5**	**71.5**	**74.3**	**75.3**	**75.4**	**84.8**	**86.1**	**84.9**	**83.8**	**82.7**	**69.9**	**43.7**	**33.6**

Table 1. Comparing with the state-of-the-art on STL-10, CelebA, CIFAR-10, CIFAR-100 and TinyImageNet datasets. For STL-10 and CelebA, we report results obtained by different convolutional layers of the backbone AlexNet (c1-5). The results report test set performance of the linear classifiers on the remaining datasets trained on the frozen layers of AlexNet models pre-trained with the pretext tasks.

4. Experiments

In this section, we perform an extensive experimental evaluation of our model on several unsupervised feature learning benchmarks and ablate our model components.

Datasets: We evaluate our model on 9 benchmark datasets: STL-10 [3], CIFAR-10 [20], CIFAR-100 [20], TinyImageNet [23], ImageNet-1K [32], Places 205 [38], CelebA [24], PASCAL VOC 2007 [6], PASCAL VOC 2012 [6]. Among these, STL-10 (10 classes, 100K unlabeled and 5K labeled images for training, 8K for test) and CIFAR-10 (10 classes, 60K images) are small scale, CIFAR-100 (100 classes, 60K images) and TinyImageNet (200 classes, 110K images) are medium scale, and ImageNet-1K (1K classes, 1.2M images) and Places 205 (205 classes, 2.4M images) are large scale classification datasets with a single label per image. CelebA (40 classes, 182K images) is a facial attribute dataset while PASCAL VOC 2007 (20 classes, 5K training, 5K test images) and PASCAL VOC 2012 (21 classes, 10K training, 10K test images) are image classification, object detection and semantic segmentation datasets where there are multiple labels per image. On STL-10, CIFAR-10, CIFAR-100, TinyImageNet, ImageNet-1K and Places 205, the evaluation metric is top-1 accuracy. On CelebA and PASCAL VOC 2007, the evaluation metric is mAP (mean average precision) and on PASCAL VOC 2012, it is mIoU (mean intersection over union).

Implementation Details: For a fair comparison with prior works, we implement our CNN model F_θ as the standard AlexNet architecture [21]. Following prior works [10, 8, 17], we remove the local response normalization layers and add batch normalization to all layers except for the final one. No other modifications to the original architecture are made. For experiments on lower resolution images (*e.g.*, STL-10), we remove the max-pooling layer after conv5 and use the default padding setting throughout the network. The auxiliary data augmentation strategies (random cropping, horizontal flipping, color jittering and blurring) are used and validated through experiments in the following sections. The size of the patch boundary is set to 2 pixels in experiments on STL-10 and CelebA. On ImageNet-1K, we use a 4 pixel boundary.

4.1. Comparing with the State-of-the-Art

Methods: To evaluate the effectiveness of our hybrid CODIAL model that follows self-supervised learning protocol, we compare it with three state-of-the-art methods on five datasets. RotNet [10] discriminating rotation angles and GlobStat [17] distinguishing rotation angles, warping and limited context inpainting are considered as discriminating methods. Decouple [8] jointly discriminates rotation angles and aligns similar features by minimizing representation distance is an instance of hybrid method.

Results: We observe in Table 1 that our method steadily improves over all the prior methods' reported results on these datasets. Specifically, the consistency over all the AlexNet layers on both the datasets shows the robustness of our model to learn low level image features to high level semantic features. On CelebA, our model particularly outperforms the prior works by a large margin as a mutual benefit of our proposed discriminative and aligning objectives. We also observe that on CelebA dataset, similar to other prior works, our model performs better in lower layers as compared to the higher layers, as lower level geometric features are more important for recognizing different facial attributes than the high level semantic features of the higher layers. On the challenging STL-10, our method respectively outperforms recently proposed Decouple and GlobStat methods by an average margins of 1.9% and 2.1%, and on CelebA, it respectively obtains a margin of 12.3% and 12.0%. This indicates the precedence of our hybrid method that concurrently discriminates and aligns features, and the mutual information based alignment strategy.

Our CODIAL surpasses the self-supervised benchmarks on CIFAR-10, CIFAR-100 and TinyImageNet datasets with a good margin, which further emphasizes the benefit of our hybrid learning approach. Particularly, we surpass GlobStat method on CIFAR-10, CIFAR-100 and TinyImageNet datasets respectively by the margins of 1.8%, 4.5% and 2.4% which shows the effectiveness of our method. We exceed Decouple respectively by the margins of 4.0%, 6.0% and 5.1% on the same datasets, as an advantage of our MI based alignment strategy, since Decouple aligns features by minimizing representation distance.

	STL-10					CelebA				
Transf.	`c1`	`c2`	`c3`	`c4`	`c5`	`c1`	`c2`	`c3`	`c4`	`c5`
cr	59.9	69.8	73.4	73.9	74.3	84.7	85.7	84.0	81.8	79.5
cr + bl	59.5	69.5	71.9	72.8	72.9	84.6	85.8	84.1	81.9	79.8
cr + co	**60.5**	**71.5**	**74.3**	**75.3**	**75.4**	84.8	**86.1**	**84.9**	**83.8**	**82.7**
cr + bl + co	59.6	71.0	73.2	74.8	75.1	**84.9**	85.9	**84.9**	83.7	82.5
Loss	`c1`	`c2`	`c3`	`c4`	`c5`	`c1`	`c2`	`c3`	`c4`	`c5`
Random	48.4	53.3	51.1	48.7	47.9	68.9	70.1	66.7	65.3	63.2
$\lambda_{\text{cls,mi}}$ = 1,0	57.8	68.7	69.2	68.1	66.4	74.5	75.5	74.0	72.6	71.5
$\lambda_{\text{cls,mi}}$ = 0,1	58.2	67.0	68.9	67.9	66.5	71.2	73.7	72.2	69.4	67.7
$\lambda_{\text{cls,mi}}$ = 1,1	**60.5**	**71.5**	**74.3**	**75.3**	**75.4**	**84.8**	**86.1**	**84.9**	**83.8**	**82.7**

Table 2. Ablating different transformations (top): We report test set performance of our pre-text model trained on different transformations. Ablating the loss weights (bottom): We report the test set performance of our final model by varying the loss weights. (STL-10 and CelebA with CNN Backbone AlexNet with conv layers (`c1-5`), cr: crop, bl: blur, co: color transformations).

4.2. Ablating Model Components

We perform our ablation study on STL-10 and CelebA datasets and analyze different design choices. For all the experiments in this section, we first pre-train our model for solving the self-supervised pretext tasks then we employ linear classifiers on top of the frozen convolutional features to classify the images from the respective datasets.

Auxiliary Image Transformations: We consider several auxiliary image transformations or augmentations, such as random horizontal flipping, random cropping, color jittering and blurring. Among them, random cropping is done by selecting a random patch of the image with an area uniformly sampled between 8% and 100% of that of the original image, and an aspect ratio logarithmically sampled between 3/4 and 4/3. This patch is then resized to a squared image of size dependent on the datasets. In case of the color jittering transform, we shift brightness, contrast, saturation, and hue of each pixel in the image by a uniformly sampled offset from the range $[0.5, 1.5]$. The order in which these shifts are performed is randomly selected for each image. Gaussian blurring is implemented with a Gaussian kernel of size 10% of that of the image and a standard deviation uniformly sampled over $[0.1, 2.0]$. In this experiment, we systematically analyze the contributions of the above mentioned image transformations on our model. For doing so, we pre-train our model with the above augmentations on the STL-10 and CelebA datasets and report the performance of linear classifiers trained on top of the frozen AlexNet layers.

From Table 2 (top) we observe that combining color jittering with other transformations improves the results across all the AlexNet layers on both the datasets, however, the difference is more in case of STL-10 as it contains natural images with diverse colors compared to CelebA dataset which mainly contains face images where only skin color has a prevalence. As a result, color jittering produces better outcome on STL-10 compared to CelebA dataset. We also observe that on STL-10, the difference of accuracy with `conv1`, `conv2`, and `conv3` features are respectively 1.0%, 2.0% and 2.4%, while on CelebA dataset, the difference of mAP values are 0.1%, 0.4% and 0.9%. For CelebA dataset, we speculate that since our self-supervised features are intend to learn high level semantic features, it does not waver much the geometric features.

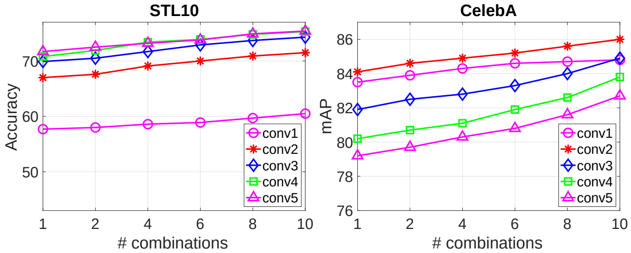

Figure 3. Plots showing the performance of five convolutional layers of our AlexNet model trained with increasing number of paired transforms, exhibit that a large number of transformed pairs is important for better performance.

Loss function: We compare the impact of two parts of our proposed loss function (eqn. (7)). For that, we pre-train our model by selecting the weights of λ_{cls} and λ_{mi} from $\{0, 1\}$ in eqn. (7) on the STL-10 and CelebA datasets and train a linear classifier on top of different frozen layers of AlexNet.

As shown in Table 2 (bottom), considering the discriminative part together with the mutual information maximization substantially boosts the performance of the method. Specifically, on STL-10 dataset, we obtain 2.7%, 2.8%, 5.1%, 7.2% and 9.0% accuracy boost and on CelebA dataset we obtain 10.3%, 10.6%, 10.9%, 11.2%, 11.2% mAP raise respectively on the `conv1`, `conv2`, `conv3`, `conv4` and `conv5` features, which is quite significant. This can be seen as a benefit of explicitly specifying which features should be close to or far from each other.

Multiple Image Transformations: By design, our model can exploit more than one pair of transformations of the same image. We design this experiment to verify whether considering more than one image pair has any benefit on the performance of the model. For that, we choose the same STL-10 and CelebA datasets, and pre-train our model with a subset of pairs uniformly selected from all the 10 ($^{K}C_2$ and $K = 5$ in our case) transformed pairs. Once trained, the different layers of our AlexNet model are evaluated on the same downstream task on the same STL-10 and CelebA datasets. Figure 3 shows the performance of different layers of AlexNet for a variant number of paired transformations. The plots shown in the figure exhibits climbing trend for all the layers as the number of paired transformations increases and for all the layers the best performances are obtained when all possible transformed pairs are considered. This indicates that a large number of transformed pairs is important for effectively maximizing mutual information and our

	ImageNet-1K					Places 205					Pascal VOC		
Model / Layer	c1	c2	c3	c4	c5	c1	c2	c3	c4	c5	cls.	det.	seg.
Places Labels	Not applicable					22.1	35.1	40.2	43.3	44.6	Not applicable		
ImageNet Labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7	79.9	59.1	48.0
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5	53.3	43.4	19.8
DeepCluster [1]	12.9	29.2	38.2	39.8	36.1	18.6	30.8	37.0	37.5	33.1	73.7	55.4	45.1
RotNet [10]	18.8	31.7	38.7	38.2	36.5	21.5	31.0	35.1	34.6	33.7	73.0	54.4	39.1
KnowTrans [29]	19.2	32.0	37.3	37.1	34.6	19.2	32.0	37.3	37.1	34.6	72.5	56.5	42.6
Decouple [8]	19.3	33.3	40.8	41.8	**44.3**	22.9	32.4	36.6	37.3	**38.6**	74.3	57.5	45.3
SpotArtifacts [16]	19.5	33.3	37.9	38.9	34.9	23.3	34.3	36.9	37.3	34.4	74.3	57.5	45.3
GlobStat [17]	20.8	34.5	40.2	43.1	41.4	24.1	33.3	37.9	39.5	37.7	74.5	56.8	44.4
CMC [34]	18.3	33.7	38.3	40.5	42.8	-	-	-	-	-	73.8	56.6	44.8
CODIAL (ours)	**22.1**	**36.2**	**41.8**	**44.3**	42.2	**25.5**	**34.6**	**38.9**	**41.2**	38.4	**75.2**	**58.3**	**45.5**

Table 3. Validation set accuracy (%) on ImageNet-1K and Places 205 datasets (left) with linear classifiers trained on frozen convolutional layers and transfer learning results for classification, detection (mAP) and segmentation (mIoU) on PASCAL VOC (right) compared to state-of-the-art feature learning methods (c1-5: convolutional layers 1-5 of AlexNet, cls: classification, det: detection, seg: segmentation).

model design constructively supports that necessity.

4.3. Unsupervised Feature Learning

In addition to RotNet [10], Decouple [8] and GlobStat [17], we compare our method with four more recently proposed methods. Among them, DeepCluster [1] iteratively clusters the features with kmeans algorithm, and uses the cluster labels as supervision to update the weights of the network; KnowTrans [29] uses clustering to boost the self-supervised knowledge transfer; SpotArtifacts [16] learns self-supervised knowledge by spotting synthetic artifacts in images; CMC [34] learns self-supervised knowledge by contrasting multiple view of the same data. We pre-train our model for 100 epochs on ImageNet-1K [32], where the images are cropped to 224×224. The pre-training was done with batch size of 256 and on 2 Titan RTX GPUs.

Linear Classification on ImageNet-1K and Places 205: Following [36, 8, 17], we train linear classifiers on top of the frozen features extracted by different convolutional layers measuring the task specific power of the learned representations, specifically the discriminative power over object class. For the experiments on the ImageNet-1K and Places 205 datasets in Table 3 (left), all the approaches use AlexNet as the backbone network and all the methods except 'ImageNet Labels', 'Places Labels' and 'Random' are pre-trained on ImageNet-1K without labels. 'ImageNet Labels' and 'Places Labels' are supervised benchmarks and are respectively trained with ImageNet-1K and Places 205 labels. All the weights of the feature extractor network are frozen and feature maps are spatially resized (with adaptive pooling) so as to have the feature dimension around 9,000.

Our learned features appear to be very robust and achieve state-of-the-art performance with all the layers from conv1 to conv4 on ImageNet-1K, particularly on conv4 features, we obtain 1.2% improvement that GlobStat, which is quite remarkable. Our result on conv5 also surpasses the recent benchmark GlobStat [17] by 0.8% and is comparable with the best result obtained by Decouple [8]. Considering the fact that the lower layers of a network usually capture low-level information like edges or contours in images and the higher layers extract abstract semantic information, our model shows diverse capacity in capturing low-level as well as high-level abstract features which can be considered as the benefit of our hybrid model. We also observe the same on the Places 205 dataset and achieve the state-of-the-art results with all layers from conv1 to conv4. Our result on conv4 in particular is the overall best among all the methods except the conv4 and conv5 layers of 'Places Labels' which is a fully supervised model. Note, that our method surpasses the performance of an AlexNet trained on ImageNet-1K using supervision on the conv1, conv3 and conv4 layers respectively by 2.8%, 0.5%, 1.8%.

Transfer Learning on PASCAL VOC: We test the transferability of the learned feature on PASCAL VOC dataset [6]. We use our unsupervised trained network F_θ as the initialization model for variant tasks, such as multi-label image classification, object detection and semantic segmentation. Performance is measured by mean average precision (mAP) for classification and detection, and by mean intersection over union (mIoU) for segmentation task. We follow the established setup of [22] for multi-label classification, the Fast-RCNN framework [11] for detection and the FCN framework [33] for semantic segmentation. For classification, we extract the conv5 features on the PASCAL VOC 2007 images and train a regularized multinomial logistic regression classifier on top of those frozen representation by minimizing the cross-entropy objective using LBFGS with ℓ_2-regularization with standard parameters. In the case of object detection, the self-supervised weights learned by our model are used to initialize the Fast-RCNN model [11] and use a multi-scale training and single-scale testing on the PASCAL VOC 2007 images. For segmen-

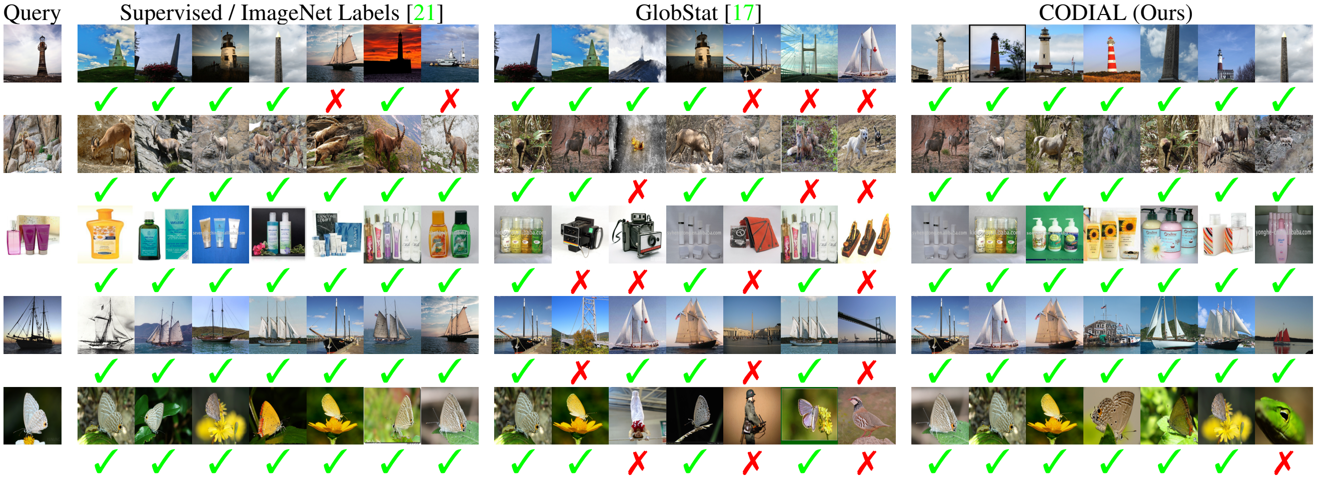

Figure 4. Comparison of nearest neighbor retrieval results. The left most column shows the query images. In each of the three consecutive columns, we show the seven nearest neighbors of the query image with features respectively learned using ImageNet labels, GlobStat [17] and our model CODIAL. Queries contain randomly selected images possessing variant characteristics. Semantically related and unrelated retrievals are respectively marked with green tick and red cross. (best viewed in color)

tation, we fine-tune the self-supervised features learned by our model using FCN [33] on PASCAL VOC 2012.

Table 3 (right) summarizes the comparison of our approach with other methods. We consistently outperform previous methods on all these three tasks, particularly on classification, detection and segmentation tasks, we respectively achieve 0.7%, 0.8% and 0.2% margin gain compared to the challenging prior arts, which is quite remarkable and can be seen as the advantage of our hybrid method that jointly solve the discriminative and aligning tasks.

4.4. Qualitative Results

Self-supervised training associates similar features to semantically similar images. In this section, we visually investigate whether our features fulfill this property. Additionally, we qualitatively compare our results with the features learned by supervised 'ImageNet Labels' model and the recent self-supervised model GlobStat [17]. For doing so, we compute the nearest neighbors of the SSL (for model proposed by GlobStat [17] and us) and SL (model trained with 'ImageNet Labels') features of `conv5` layers of AlexNet on the validation set of ImageNet-1K. For our model, we obtain features from the 4,096 dimensional vector outputted by the feature extractor network F_θ. We use cosine similarity to calculate the distance between features.

The images are arranged from left to right in order of increasing distance in Figure 4. We observe that all the models are able to capture semantic information in images for some queries. In some cases, our retrievals are quite similar to the ones returned by GlobStat [17], however, many images retrieved by [17] are unrelated to the query. This is likely because GlobStat [17] focuses more on the texture and shape of object and is less discriminative towards different instances. As a result, it retrieves many instances based on the background information, such as in rows 1 and 4, some wrongly retrieved images are found based on the texture of sky. In row 2, background and shape of the query animal influence the wrong retrieval. Also in row 3, some wrong examples are retrieved based on the white background, we speculate that happens because of its ability to focus on texture information. On the contrary, our model can return more semantically similar images for these queries, which confirms our model's discriminative ability on instance level as a benefit to the maximization of mutual information.

5. Conclusion

We presented CODIAL, a self-supervised visual representation learning method by jointly solving discriminating and aligning tasks. The discriminative task involves recognizing image transformations, such as rotation angle, warping, that needs learning global statistics of the image, whereas the alignment is done by maximizing the mutual information between the transformed version of the same image. This principle explicitly describes which features to be closed together or separated. We present the efficacy of our proposal through substantial experiments on nine self-supervised and transfer learning benchmark datasets where our model consistently outperforms the existing methods in various settings and tasks.

Acknowledgments

This work has been partially funded by the ERC starting grant DEXIM (grant agreement no. 853489) and DFG-EXC-Nummer 2064/1-Projektnummer 390727645.

References

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, 2018. 2, 7

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020. 1, 2

[3] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *AISTATS*, 2011. 5

[4] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015. 1, 2, 3

[5] Carl Doersch and Andrew Zisserman. Multi-task Self-Supervised Visual Learning. In *ICCV*, 2017. 1

[6] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge: A Retrospective. *IJCV*, 2015. 5, 7

[7] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*, 2020. 1, 2, 4

[8] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-Supervised Representation Learning by Rotation Feature Decoupling. In *CVPR*, 2019. 1, 2, 3, 5, 7

[9] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to Classify Images without Labels. In *ECCV*, 2020. 2

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*, 2018. 1, 2, 3, 5, 7

[11] Ross Girshick. Fast R-CNN. *ICCV*, 2015. 7

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv*, 2020. 1, 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 2019. 2

[14] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 2, 4

[15] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S M Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. In *ICML*, 2020. 1, 2

[16] Simon Jenni and Paolo Favaro. Self-Supervised Feature Learning by Learning to Spot Artifacts. In *CVPR*, 2018. 1, 2, 3, 7

[17] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics. In *CVPR*, 2020. 1, 2, 3, 5, 7, 8

[18] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *ICCV*, 2019. 1

[19] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE TPAMI*, 2020. 1

[20] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. 5

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 5, 8

[22] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent Initializations of Convolutional Neural Networks. In *ICLR*, 2016. 7

[23] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Tiny ImageNet Visual Recognition Challenge. Dataset, Stanford University, 2015. 5

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[25] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. In *CVPR*, 2020. 1, 2

[26] T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. In *CVPR*, 2018. 2

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV*, 2016. 2

[28] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation Learning by Learning to Count. In *ICCV*, 2017. 2

[29] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting Self-Supervised Learning via Knowledge Transfer. In *CVPR*, 2018. 2, 7

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. In *arXiv*, 2018. 1, 2

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In *CVPR*, 2016. 2

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2014. 5, 7

[33] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE TPAMI*, 2016. 7, 8

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In *ECCV*, 2020. 1, 2, 7

[35] Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. In *ITW*, 2015. 2, 4

[36] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful Image Colorization. In *ECCV*, 2016. 2, 7

[37] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. In *CVPR*, 2017. 2

[38] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning Deep Features for Scene Recognition using Places Database. In *NIPS*, 2014. 5