# Residual Contrastive Learning for Image Reconstruction: Learning Transferable Representations from Noisy Images

**Nanqing Dong**[1*†], **Matteo Maggioni**[2], **Yongxin Yang**[2], **Eduardo Pérez-Pellitero**[2], **Ales Leonardis**[2], **Steven McDonagh**[2*]

[1]Department of Computer Science, University of Oxford
[2]Huawei Noah's Ark Lab

## Abstract

This paper is concerned with contrastive learning (CL) for low-level image restoration and enhancement tasks. We propose a new label-efficient learning paradigm based on residuals, *residual contrastive learning* (RCL), and derive an unsupervised visual representation learning framework, suitable for low-level vision tasks with noisy inputs. While supervised image reconstruction aims to minimize residual terms directly, RCL alternatively builds a connection between residuals and CL by defining a novel instance discrimination pretext task, using residuals as the discriminative feature. Our formulation mitigates the severe task misalignment between instance discrimination pretext tasks and downstream image reconstruction tasks, present in existing CL frameworks. Experimentally, we find that RCL can learn robust and transferable representations that improve the performance of various downstream tasks, such as denoising and super resolution, in comparison with recent self-supervised methods designed specifically for noisy inputs. Additionally, our unsupervised pre-training can significantly reduce annotation costs whilst maintaining performance competitive with fully-supervised image reconstruction.

## 1 Introduction

Fueled by the advances of self-supervised learning[1] (SSL), large-scale unsupervised pre-training followed by fine-tuning on small amounts of annotated data has become a popular label-efficient learning paradigm. A standard example involves firstly unsupervised visual representation learning (UVRL) using ImageNet [Deng *et al.*, 2009]. The learned representations can then be transferred to downstream tasks, reducing the number of labels required and yet achieving strong performance, competitive with supervised pre-training [He *et al.*, 2020]. Self-supervised strategies therefore have the potential to reduce labeling costs and this becomes especially pertinent for dense prediction tasks where human annotation is time-consuming and often very expensive.

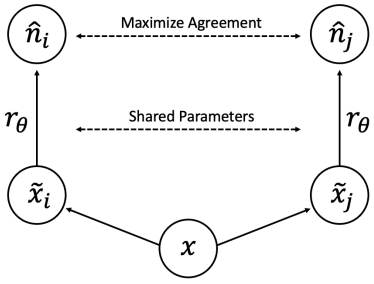


Figure 1: RCL for low-level visual representations with noisy inputs. $\boldsymbol{x}$ is a noisy image. $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ are two random crops from the same $\boldsymbol{x}$ (a positive pair). $r_\theta(\cdot)$ is a residual function, defined in Eq. 2. $\hat{\boldsymbol{n}}_i$ and $\hat{\boldsymbol{n}}_j$ are two corresponding residual tensors.

Further, an unsupervised pre-training method, capable of learning representations transferable across various downstream tasks, is desirable as this enables efficiency in terms of both computation and labeling costs. Without this property it quickly becomes impractical to label large datasets for multiple tasks of interest or alternatively design appropriate individual pretext tasks for each downstream application.

One such SSL strategy capable of learning transferable representations [Wang and Isola, 2020] is *contrastive learning* (CL) [Chen *et al.*, 2020; He *et al.*, 2020; Chuang *et al.*, 2020; Li *et al.*, 2021]. CL is based on an instance discrimination pretext task [Wu *et al.*, 2018], where the learning goal is to maximize the mutual information of two views of the same image, pulling two augmented image views together in the feature space, whilst pushing apart representations of different images [Oord *et al.*, 2018]. While CL has been shown to provide comparable performance with, and even improve upon, supervised pre-training with respect to various high-level downstream tasks [Chen *et al.*, 2020; He *et al.*, 2020], contemporary strategies expose two limitations for low-level image restoration and enhancement tasks.

Firstly, recent studies have shown that when the pretext and downstream tasks are not closely correlated, improving pretext task performance cannot guarantee downstream task improvement [Ericsson *et al.*, 2021]. This phenomenon is

*Contact Authors

†Part of this work was done at Huawei Noah's Ark Lab.

[1]We use the terms "self-supervised learning" and "self-supervised representation learning" interchangeably in this work.

known as *task misalignment* [Dong *et al.*, 2021]. However, existing CL frameworks are mainly designed for high-level semantic understanding tasks, leaving the potential of CL in conjunction with low-level vision domains currently under-explored. Empirically, we find that the task misalignment can severely impair the representation learning performance of existing CL frameworks for low-level downstream tasks.

Secondly, images utilized for CL pre-training are typically assumed to be noise free, yet input to image enhancement and restoration tasks commonly contain additive noise. This is compounded by commonly adopted CL data augmentation policies [Chen *et al.*, 2020] that affect the data distribution and encourage the learning of invariances less relevant for image reconstruction tasks. Alternative SSL approaches have however been designed specifically for noisy images. Although such methods have achieved promising results for the denoising task specifically [Batson and Royer, 2019], we find that they are less well equipped to efficiently learn transferable representations when the downstream data distributions change (*i.e.* for additional low-level vision tasks).

Motivated by these considerations, our work aims to answer an under-explored question: *how can CL be used to learn transferable representations for low-level vision tasks, from noisy images?*

We start by recalling standard supervised learning (SL). Let ($x$, $y$) define an input and target image reconstruction respectively (*e.g.* a noisy and noise-free image pair in the denoising literature), the loss can then be formulated as $\|y - f_\theta(x)\|$, where $f_\theta(\cdot)$ is the model of interest with parameters $\theta$. This canonical use of paired data provides supervised models with a useful signal however obtaining ground truth data for real-world image enhancement and restoration tasks that necessitate dense prediction may require complex and often constraining procedures, *i.e.* $y$ is often unavailable due to annotation costs. Removing the requirement of a noise-free image $y$, and instead minimizing $\|x - f_\theta(x)\|$, can be seen to provide a trivial solution where $f_\theta(\cdot)$ is an identity mapping. Various SSL efforts therefore instead propose to minimize more useful objectives of the form $\|\tilde{x} - f_\theta(x)\|$, where $\tilde{x}$ constitutes *e.g.* a second noisy variant of $x$ [Lehtinen *et al.*, 2018; Batson and Royer, 2019; Ehret *et al.*, 2019]. We observe that, without the norm operator, $x - f_\theta(x)$ can be regarded as a *residual* term. In statistics and optimization, a *residual* denotes the difference between observed and estimated values of interest. In the domain of deep learning residuals commonly take the form $r(x) = f_\theta(x) - x$, where $x$ is the input, $f_\theta(x)$ is the output, and $r(\cdot)$ is the residual function [He *et al.*, 2016].

Following this formulation, we propose *residual contrastive learning* (RCL), a residual-based SSL framework for noisy images (illustrated in Fig. 1). We bridge a methodological gap between SSL on visual signals with additive noise and unsupervised residual learning via CL. We conjecture that residuals can be effectively used as a discriminative feature for CL based on the fact that additive image noise is signal-dependent [Hasinoff *et al.*, 2010]. We propose a *residual contrastive loss*, which leverages the earth mover's distance (EMD) to measure the similarity between two residual tensors with the same shape (*c.f.* cosine similarity applied to two feature vectors [Chen *et al.*, 2020]). By leveraging signal-dependent noise as an appropriate discriminative feature based on the prior knowledge in image processing, in tandem with the representation learning ability of CL, RCL is expected to learn transferable representations amenable to downstream image reconstruction tasks.

Similar to previous CL studies that alternatively consider representation learning for high-level vision tasks [Chen *et al.*, 2020], we adopt a *proxy evaluation* protocol that uses the performance of proxy supervised downstream tasks to measure the quality of the representations learned during unsupervised pre-training. We establish a set of benchmark datasets and downstream tasks towards systematically evaluating RCL against both (i) recent CL methods that focus on dense prediction [O. Pinheiro *et al.*, 2020; Wang *et al.*, 2021a; Xie *et al.*, 2021] and (ii) strong SSL methods designed specifically for noisy inputs [Lehtinen *et al.*, 2018; Batson and Royer, 2019; Ehret *et al.*, 2019]. We observe that representations learned by RCL consistently outperform the baselines and exhibit strong generalization ability in multiple downstream tasks; namely denoising, super resolution and demosaicing. Finally, we report that a learning paradigm involving pre-training on unlabeled data using RCL, followed by fine-tuning on small labeled data with SL, can achieve performance competitive with fully-supervised baselines. In summary, our contributions are as follows:

1. We provide the first formulation of an instance discrimination pretext task based on residuals.
2. We propose RCL, a novel framework that can learn transferable representations from only noisy inputs. To the best of our knowledge, this constitutes the first study of CL on noisy images for low-level image reconstruction tasks.
3. Our empirical results show that RCL learns robust representations from noisy images without paired ground truth, and unsupervised pre-training with RCL can significantly reduce the annotation cost in comparison with fully-supervised alternatives.

## 2 Related Work

The recent renaissance of CL has been driven by the successes of UVRL on ImageNet [Chen *et al.*, 2020; He *et al.*, 2020]. In the case of image recognition, the objective of both the instance discrimination pretext task and corresponding downstream task are highly correlated and intuitively, so is resulting performance. However, for tasks such as object detection and others involving dense prediction, correlation still exists yet is found to be weaker than in the case of recognition [Ericsson *et al.*, 2021]. To mitigate such task misalignment, several state-of-the-art (SOTA) CL frameworks, designed for dense prediction tasks, have been proposed, *e.g.* VADeR [O. Pinheiro *et al.*, 2020], DenseCL [Wang *et al.*, 2021a], and PixContrast [Xie *et al.*, 2021]. These approaches perform pixel-wise CL and train encoder-decoder networks that directly enable dense prediction. However, in contrast to RCL, these SOTA approaches are designed and evaluated for *semantic* understanding tasks, *i.e.* task misalignment still exists when the downstream task is related to image reconstruction. A further significant difference between this work

and those highlighted is that we purposefully abstain from relying on data augmentation. We find that augmentation can alter the original data noise distribution and potentially provides a signal that leads to learning invariances, undesirable for our target low-level tasks.

There have additionally been recent applications of CL to *specific* low-level vision tasks, *e.g.* dehazing [Wu *et al.*, 2021] and super resolution [Wang *et al.*, 2021b]. In these studies, CL is considered as a regularization technique to provide an end-to-end solution for the specific task. In contrast, our proposed approach attempts to provide a step towards a *universal* UVRL framework for unprocessed images with natural noise, instead of for task-dependent applications. Thus, RCL could be used as a pre-training step for these downstream tasks.

A recent SSL study that, similar to our work, makes use of an EMD metric is *self*-EMD [Liu *et al.*, 2020]. In contrast to our proposed approach, their method formulates object detection as an optimal transport problem whereas we directly measure the similarity between two distributions without requiring an iterative procedure, commonly induced by optimal transport problems that make use of the Sinkhorn-Knopp algorithm.

# 3 Residual Contrastive Learning

## 3.1 Preliminary

A widely adopted contrastive loss, InfoNCE [Oord *et al.*, 2018], is formulated as:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(z_q, z_0)/\tau)}{\sum_{i=0}^{N} \exp(\text{sim}(z_q, z_i)/\tau)} \tag{1}$$

where $z$ denotes the feature vector extracted from an image patch of interest, $\tau$ is a temperature parameter, and $\text{sim}(\cdot,\cdot)$ is the cosine similarity function. Firstly image patches are encoded into feature vectors via an encoder and then $\text{sim}(\cdot,\cdot)$ can be used to measure the similarity between these representations. $(z_q, z_0)$ is a positive pair such that two augmented views are taken from the same image; and $(z_q, z_{i>0})$ is a negative pair, where two patches are taken from different images.

## 3.2 Problem Formulation

We denote $\boldsymbol{x}$ as a noisy image signal with clean image signal $\boldsymbol{y}$ and additive noise $\boldsymbol{n}$. The relation[2] between the tuple $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{n})$ can be formed as $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{n}$. The noise element $\boldsymbol{n}$ follows an unknown signal-dependent distribution.

For a low-level vision task under SL, a training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_\mathcal{S}}$ is given, with $N_\mathcal{S}$ training examples. Let $f_\theta$ denote a model of interest which takes $\boldsymbol{x}$ as input. The optimization goal is then to minimize $\|f_\theta(\boldsymbol{x}) - \boldsymbol{y}\|_p$, for optimal model weights $\theta$ where $\|\cdot\|_p$ denotes the $p$-norm.

The problem setting of interest in this work is UVRL, which involves an unlabelled training set. We alternatively consider $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^{N_\mathcal{S}}$ and the goal is to learn representations (*i.e.* optimize model weights $\theta$) for downstream tasks with access to the noisy image signals only.

[2]For simplicity, we assume $\boldsymbol{x}$ and $\boldsymbol{y}$ take a common image format, *e.g.* RGB or RAW.

## 3.3 Residual-Based Instance Discrimination

A key contribution of this study consists of the formulation of our *residual*-based instance discrimination pretext task. We prime this by noting that supervised residual learning has led to success in many low-level vision tasks [Zhang *et al.*, 2017; Li *et al.*, 2018]. The residual tensor for $\boldsymbol{x}$ is defined as

$$\hat{\boldsymbol{n}}(\boldsymbol{x}) = r_\theta(\boldsymbol{x}) = \boldsymbol{x} - f_\theta(\boldsymbol{x}). \tag{2}$$

We use the residual tensors as the discriminative input for CL. We are motivated by the empirical observation that on average; *the noise distributions associated with two image crops, extracted from the same image, have detectably smaller divergence than noise distributions pertaining to crops extracted from different images*. This observation constitutes a natural extension of the signal-dependency assumption. Further, for natural images, the noise distributions of two crops originating from the same instance may also possess high correlation due to potential self-similarities [Batson and Royer, 2019], with similar structures appearing at different locations and scales in the same image.

## 3.4 Residual Contrastive Loss

We now formally introduce the proposed *residual contrastive loss*. Note, the $\text{sim}(\cdot,\cdot)$ function in Eq. 1 implicitly imposes two constraints: (i) the input $z$ is required to take the form of a normalized vector; and (ii) an element-wise correspondence between two feature vectors is required in the feature space. To realize a residual contrastive loss suitable for dense prediction tasks associated with low-level vision, we relax these constraints by replacing the cosine similarity $\text{sim}(\cdot,\cdot)$ with a negative distance function. The original contrastive loss (Eq. 1) can then be reformulated as

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(-\text{d}(\hat{\boldsymbol{n}}(\boldsymbol{x}_q), \hat{\boldsymbol{n}}(\boldsymbol{x}_0))/\tau)}{\sum_{i=0}^{N} \exp(-\text{d}(\hat{\boldsymbol{n}}(\boldsymbol{x}_q), \hat{\boldsymbol{n}}(\boldsymbol{x}_i))/\tau)} \tag{3}$$

where $\tau$ is a temperature parameter [Chen *et al.*, 2020] and $\text{d}(\cdot,\cdot)$ is a non-negative statistical metric measuring the divergence between two probability distributions, such that larger metric values indicate larger divergence.

**Distance Function** We further note that, unlike cosine similarity, $\text{d}(\cdot,\cdot)$ should not assume a pair-wise relationship between two samples, as the noise distribution is independent of the pixel location. Valid distance measures $\text{d}(\cdot,\cdot)$ should also possess desirable properties such as ease of computation and differentiability, towards enabling efficient end-to-end training. Common information theoretic measures that require density estimation (*e.g. Kullback–Leibler* divergence) do not meet the above requirements. In this work, we choose the *earth mover's distance* (EMD). Let $(\hat{\boldsymbol{n}}(\boldsymbol{x}_p), \hat{\boldsymbol{n}}(\boldsymbol{x}_q))$ be two residual tensors, we then have

$$\begin{aligned}&\text{EMD}(\hat{\boldsymbol{n}}(\boldsymbol{x})_p, \hat{\boldsymbol{n}}(\boldsymbol{x})_q) = \\ &\inf_{\gamma \in \Pi(P_p, P_q)} \mathbb{E}_{(\hat{\boldsymbol{n}}(\boldsymbol{x})_p, \hat{\boldsymbol{n}}(\boldsymbol{x})_q) \sim \gamma}[\|\hat{\boldsymbol{n}}(\boldsymbol{x})_p - \hat{\boldsymbol{n}}(\boldsymbol{x})_q\|],\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{n}}(\boldsymbol{x})_p \sim P_p$, $\hat{\boldsymbol{n}}(\boldsymbol{x})_q \sim P_q$, and $\Pi(\cdot,\cdot)$ denotes the joint distribution.

**Training** For computational simplicity, we define a positive pair $(\hat{\boldsymbol{n}}(\boldsymbol{x}_q), \hat{\boldsymbol{n}}(\boldsymbol{x}_0))$ as two overlapping image patches

**Algorithm 1** Batch-wise training of *residual contrastive loss*

1: Sample a batch of $N+1$ images. ▷ Sample $N+1$ positive pairs
2: Sample two positive patches for each image.
3: Generate $\hat{\boldsymbol{n}}(\boldsymbol{x})$ for each of $2N+2$ patches. ▷ Eq. 2
4: **for** $j = 1, 2, \ldots, N+1$ **do**
5: Take the $j^{\text{th}}$ pair as the positive pair $(\hat{\boldsymbol{n}}(\boldsymbol{x}_q), \hat{\boldsymbol{n}}(\boldsymbol{x}_0))$.
6: Take the $2^{\text{nd}}$ patch of each of the other $N$ pairs as $\hat{\boldsymbol{n}}(\boldsymbol{x}_{i>1})$.
7: Compute $\mathcal{L}_{\text{contrast}}$ for the $j^{\text{th}}$ positive pair. ▷ Eq. 3
8: Sum up $\mathcal{L}_{\text{contrast}}$ for a batch of $N+1$ images as the batch-wise *residual contrastive loss*.

$(\boldsymbol{x}_q, \boldsymbol{x}_0)$ cropped from the same instance and a negative pair $(\hat{\boldsymbol{n}}(\boldsymbol{x}_q), \hat{\boldsymbol{n}}(\boldsymbol{x}_{i>1}))$ as two image patches $(\boldsymbol{x}_q, \boldsymbol{x}_{i>1})$ cropped from two different instances. The batch-wise training details of the residual contrastive loss are illustrated in Algorithm 1.

### 3.5 Optimization

While Eq. 3 enables UVRL, this gives rise to a further question: as $f_\theta$ could represent an arbitrary function that satisfies Eq. 2, the representations learned by RCL may not be meaningful for the downstream tasks of interest. CL works well for high-level visual representations because the pretext tasks and downstream tasks both involve discrimination of visual objects. Similarly, we require to build such a connection between RCL and low-level vision tasks.

The performance of low-level image reconstruction tasks is known to be sensitive to pixel-level intensities. This offers a simple solution: inclusion of the term $\|\boldsymbol{x} - f_\theta(\boldsymbol{x})\|$ as a regularizer. Note that minimizing $\|\boldsymbol{x} - f_\theta(\boldsymbol{x})\|$ alone (*i.e.* without Eq. 3) could lead to the trivial solution of an identity mapping. This issue can be mitigated through the introduction of non-linearities to both terms. Inspired by this strategy, we leverage the basic concept of the *perceptual loss* [Johnson *et al.*, 2016] and define a *consistency* loss term as

$$\mathcal{L}_{\text{consistency}} = \|\phi(g_e(\boldsymbol{x})) - \phi(g_e(f_\theta(\boldsymbol{x})))\|_2^2, \tag{5}$$

where $\phi(\cdot)$ represents the features extracted from a pre-trained encoder $g_e$. Note, $g_e$ could either be pre-trained in a self-supervised fashion using the unlabeled noisy inputs [He *et al.*, 2020] or acquired from existing pre-trained weights (*e.g.* from ImageNet [Deng *et al.*, 2009]). We utilize the assumption that the noisy input image and the reconstructed output image should convey similar semantic information, *i.e.* the noise should not drastically change the semantic content of the image. The final training objective is then the sum of the two introduced losses:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{consistency}}, \tag{6}$$

where $\alpha$ is a weighting parameter chosen empirically.

## 4 Experimental Setup

We introduce here experimental protocols, datasets and implementation details.

**Simulation** We aim to evaluate the generalization ability of the learned representations across different downstream tasks. However we note that real-world multi-task datasets, pertaining exclusively to low-level vision tasks, are currently scarce in the literature. Thus, to empirically validate the idea of CL with residuals, we firstly establish a set of benchmark datasets based on synthetic signal-dependent noise. To simulate such signal-dependent noise, we generate synthetic heteroscedastic Gaussian noise based on a noise level function (NLF) model ( $\boldsymbol{n} \sim \mathcal{N}(0, \lambda_{\text{shot}}\boldsymbol{x} + \lambda_{\text{read}})$). We use different $(\lambda_{\text{shot}}, \lambda_{\text{read}})$ to model different cameras and acquisition settings. The parameters $(\lambda_{\text{shot}}, \lambda_{\text{read}})$ are randomly sampled to ensure the overall noise variance level $\sigma^2$, of each image, falls in a reasonable range for the data used in our experiments and we set $\sigma \in [0, 20]$ following [Gharbi *et al.*, 2016]. We therefore consider each image to have a noise distribution with approximately *unique* parameters. From the perspective of the dataset $\mathcal{S}$, there is therefore an approximate one-to-one mapping between $(\lambda_{\text{shot}}, \lambda_{\text{read}})$ and each image. This simulation model is utilized to evaluate the robustness of SSL methods.

**Benchmark Datasets** In order to simulate large-scale unlabeled training data with signal-dependent noise, we consider three large-scale public datasets, namely, the MIT Demosaicing dataset [Gharbi *et al.*, 2016] (MIT), the Stanford Taskonomy dataset [Zamir *et al.*, 2018] (Stanford), and the PASCAL VOC dataset [Everingham *et al.*, 2010] (VOC). We generate noisy images by adding synthetic noise. The datasets are split into training and test sets.

**Proxy Evaluation** CL aims to learn strong representations for downstream tasks, *i.e.* pre-training of $f_\theta$ instead of solving each problem directly. In this work, we therefore test the generalization ability of the learned representations [Zhang *et al.*, 2016]. Following previous studies on CL for high-level vision tasks [He *et al.*, 2020; Chen *et al.*, 2020; Chuang *et al.*, 2020], we adopt a *proxy evaluation protocol*. Concretely, we fine-tune the learned representations on downstream tasks with a small amount of annotated data, under SL. We report the performance of the downstream tasks as the *proxy performance* for SSL. In this way, we can systematically evaluate the generalization and transferablility of representations learned under different SSL frameworks. Following the *linear classification protocol* [He *et al.*, 2020], first the weights of a network $f_\theta$ are pre-trained using an unlabeled training set, and then all weights except those in the last layer are frozen. The pre-trained last layer is then replaced with a randomly initiated task-dependent layer for the downstream task. The new last layer is then fine-tuned with the labeled training set and evaluated on the task test set. Note, under proxy evaluation, the representations of the intermediate layers are fixed. The reported numerical results are used to indirectly reflect the quality of fixed representations, thus this is a *proxy* evaluation. We highlight that this evaluation differs from common low-level vision task evaluation protocols, where end-to-end solutions are directly compared (without fine-tuning).

**Evaluation Metrics** We consider two common image reconstruction metrics for the proxy evaluation, *peak signal-to-noise ratio* (PSNR) and *structure similarity index measure* (SSIM). We repeat experiments over five trials and report mean results. We denote the performance of supervised pre-training as an *Oracle*.

**Implementation** Theoretically, $f_\theta$ may constitute any model capable of performing dense prediction tasks. In the following section, we will show that the representations learned by

| Method | MIT | | Stanford | | VOC | |
|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| VADeR | 14.63 | 0.1088 | 17.54 | 0.1601 | 16.33 | 0.1573 |
| DenseCL | 13.78 | 0.0910 | 16.46 | 0.1527 | 15.33 | 0.1373 |
| *PixContrast* | 14.77 | 0.1101 | 17.61 | 0.1610 | 16.42 | 0.1585 |
| VADeR+ | 19.63 | 0.4183 | 21.58 | 0.4961 | 20.75 | 0.4737 |
| DenseCL+ | 18.87 | 0.3998 | 20.58 | 0.4705 | 20.21 | 0.4647 |
| *PixContrast+* | 19.87 | 0.4121 | 21.41 | 0.4899 | 20.82 | 0.4858 |
| N2N | 28.66 | 0.8614 | 34.14 | 0.8699 | 30.91 | 0.8272 |
| N2S | 28.16 | 0.8373 | 34.04 | 0.8640 | 30.71 | 0.8256 |
| RCL-BD | 28.83 | 0.8871 | 34.75 | 0.8618 | 31.29 | 0.8274 |
| RCL-MMD | 28.68 | 0.8864 | 34.87 | 0.8687 | 31.53 | 0.8316 |
| RCL-EMD | **29.54** | **0.8908** | **35.43** | **0.8783** | **31.39** | **0.8330** |
| *Oracle* | 31.26 | 0.9187 | 38.25 | 0.9422 | 33.65 | 0.9038 |

Table 1: Proxy evaluation of representation learning using denoising as the downstream task.

| Method | SR | | JDenSR | |
|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM |
| N2N | 31.61 | 0.8730 | 27.90 | 0.7860 |
| N2S | 31.11 | 0.8699 | 27.80 | 0.7831 |
| RCL-BD | 38.89 | 0.9654 | 32.10 | 0.8046 |
| RCL-MMD | 38.31 | 0.9634 | 31.95 | 0.8068 |
| RCL-EMD | **39.01** | **0.9658** | **32.63** | **0.8214** |
| SL (Den) | 34.18 | 0.9118 | 32.89 | 0.8353 |
| *Oracle* | 38.93 | 0.9603 | 35.98 | 0.9175 |

Table 2: Proxy evaluation of representation learning using SR and JDenSR as the downstream task on the Stanford dataset.

$f_\theta$ can be successfully applied to various downstream image reconstruction tasks: denoising, demosaicing and super resolution. Following [Zamir *et al.*, 2018], we utilize a generic network backbone; U-Net [Ronneberger *et al.*, 2015] to instantiate $f_\theta$. We additionally use a ResNet50 [He *et al.*, 2016], pre-trained on ImageNet [Deng *et al.*, 2009] for the fixed feature extractor $g_e$. To instantiate Eq. 3, we follow [Chen *et al.*, 2020] in defining temperature $\tau$ values and use a batch size of 64. We use a weighting parameter $\alpha{=}10^{-3}$ in the unsupervised pre-training phase and an $L_1$ loss for the supervised fine-tuning in the evaluation phase. We use an Adam [Kingma and Ba, 2015] optimizer with $\beta_1{=}0.9$, $\beta_2{=}0.999$, and $\epsilon{=}10^{-7}$, and a fixed learning rate $10^{-3}$. The minimal image crop size is $128{\times}128$. All models are implemented in PyTorch on a NVIDIA Tesla V100 GPU.

## 5 Experiments

**Baselines** To validate the empirical considerations presented in Sec. 1, we select two sets of baselines SSL methods. We include three SOTA CL frameworks for high-level vision tasks to validate our hypothesis that there exists a task misalignment between semantic understanding tasks and image restoration tasks. We consider VADeR [O. Pinheiro *et al.*, 2020], DenseCL [Wang *et al.*, 2021a], and PixContrast [Xie *et al.*, 2021]. These three baselines apply CL at a pixel-level in the feature space, thus can train an encoder-decoder network directly for dense prediction tasks. However, in contrast to RCL, these methods are designed for semantic understanding tasks, *i.e.* task misalignment still exists. We use a consistent U-Net backbone for all three methods, where the number of output channels of the last layer is set to three (for RGB images). The pre-training and fine-tuning procedures follow Sec. 4, inline with our RCL. For a fair comparison, we also report the performance of three methods trained with additional $\mathcal{L}_{\text{consistency}}$ (Eq. 5), denoted as VADeR+, DenseCL+, and PixContrast+. We also consider two seminal SSL baselines that are designed for noisy images, namely *noise2noise* (N2N) [Lehtinen *et al.*, 2018] and *noise2self* (N2S) [Batson and Royer, 2019], for UVRL on image reconstruction tasks. For N2N, we generate paired noisy RGB images with the same random parameters ($\lambda_{\text{shot}}$, $\lambda_{\text{read}}$). Note that N2N and N2S both utilize a formulation of $\|\tilde{\boldsymbol{x}} - f_\theta(\boldsymbol{x})\|$, where $\tilde{\boldsymbol{x}}$ is a noisy observation of $\boldsymbol{x}$.

**Denoising** We instantiate denoising as the first downstream task and report representation learning results in Table 1. All three CL frameworks, designed for high-level tasks, produce results much weaker than SSL methods designed for specific low-level vision tasks, with or without $\mathcal{L}_{\text{consistency}}$. We emphasize that this is due to the highlighted severe task misalignment issue. Thus these methods are omitted in the following discussion. Note, during proxy evaluation, the pre-training set and testing set do not overlap. RCL shows competitive representation learning performance in comparison with N2N and N2S, which are reported to achieve reasonable performance in blind denoising tasks. In addition to EMD, we consider two alternative distance functions *Bhattacharyya distance* (BD) and *maximum mean discrepancy* (MMD). EMD showed more robust performance than BD and MMD in denoising and two other downstream tasks (below), we thus select EMD as our default metric for remaining experiments.

**Super Resolution** We further explore the generalization ability of our learned representations to low-level vision tasks that are markedly distinct from denoising. Super resolution (SR) constitutes such a task. As each image in the Stanford dataset has two resolutions ($512{\times}512$ and $1024{\times}1024$), we define the higher resolution image as the upsampled ground truth in order to provide a simple proof of concept SR task. Following the proxy evaluation protocols introduced previously, results are presented in Table 2 (left). We observe that RCL outperforms N2N and N2S by a large margin. By comparing Table 1 with Table 2, we find the performance gap between RCL and N2N / N2S becomes larger, *i.e.* N2N and N2S tend to learn less meaningful representations for disparate downstream tasks where the gap between them and respective pretext tasks grow, in the investigated setting. This phenomenon has also been discussed in [Zhang *et al.*, 2016], where a task-dependent colorization-based SSL shows limited classification performance. In comparison, RCL can result in representations that exhibit stronger generalisation ability.

**Joint Denoising and Super Resolution** As a natural extension to independent denoising and SR tasks, we consider a joint denoising and super resolution downstream task (JDenSR), which has two sub-tasks and can further demonstrate the versatility of the learned representations. The results are presented in Table 2 (right). Again, RCL outperforms the baseline SSL frameworks by a large margin. We note that the objective for image reconstruction tasks is typically to estimate (minimize) residuals. This is a similar setup to denoising but with the difference that the residual might have a different distribution. We hypothesize that RCL is

| # Labels | SL | | RCL + SL | |
|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM |
| 0 | - | - | 22.62 | 0.7989 |
| 10 | 20.74 | 0.7299 | 28.20 | 0.8834 |
| $10^2$ | 27.19 | 0.8734 | 30.24 | 0.9028 |
| $10^3$ | 31.31 | 0.9184 | 32.09 | 0.9280 |
| $10^4$ | 33.41 | 0.9437 | 33.85 | 0.9514 |

Table 3: Standard SL (left) and RCL pre-training with SL fine-tuning (right), evaluated with denoising on the VOC dataset. # Labels denotes the number of labeled data available for SL.

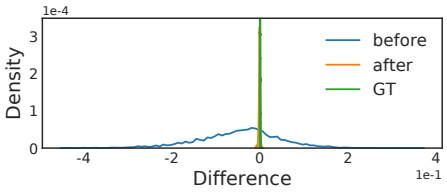


Figure 2: Comparison of residual contrastive loss (illustrated with EMD) before and after training. The density plot depicts the pair-wise differences of EMD between negative pairs minus EMD between positive pairs.

more robust to this change of distribution.

**Transferability: Supervised Pre-Training *vs*. Unsupervised Pre-Training** In addition to unsupervised pre-training, we report the performance of supervised pre-training by denoising in the "SL (Den)" row of Table 2. We learn representations by applying SL to the denoising task, defined in Table 1. We then fine-tune to the alternative downstream tasks in a fashion identical to the considered SSL methods. We note that interestingly, RCL is able to outperform "SL (Den)" for the SR task and also RCL(-EMD) achieves higher performance than the *Oracle* in Table 2. This unintuitive phenomenon, that unsupervised pre-training can improve performance over supervised pre-training, has been recently corroborated in CL studies that consider high-level vision tasks [Wang and Isola, 2020]. While supervised pre-training tends to learn task-dependent representations, the representations learned by CL are more informative. In Table 2, the "SL (Den)" row, pertaining to JDenSR results, exhibits strong performance, and a marginal advantage over RCL, which may be explained by the fact that JDenSR can be considered closely related to a pure denoising task.

**RCL *vs*. SL** To further quantify the performance and labelling-cost trade-off, we perform a sensitivity study. We train a U-Net in a supervised fashion (SL) for the denoising task using VOC data and compare this with RCL(-EMD) under various magnitudes of available training labels. We re-use the *same* random seeds for both methods. In the first row of Table 3, we report the performance of RCL by directly applying the representations pre-trained on the unlabelled training set, on the test set. In the remaining Table 3 rows, it can be observed that the performance gain obtained by pre-training with RCL grows larger as SL suffers more from label scarcity.

**Label-Efficient Learning** Our sensitivity study affords initial evidence towards answering the questions: *can RCL help SL?* and, if so, *when can RCL help?* We fine-tuned the U-Net, pre-trained by RCL(-EMD) on the entire training set, with additional paired RGB training images, as above. Pre-training with RCL consistently improves the performance of standard SL. In cases where labelled data are rare, expensive to collect or curate, such pre-training may be able to offer significant improvement (*e.g.* $+7.46dB$ with only ten labels in Table 3). We observe that fine-tuning with $500$ and $4000$ labels achieves similar performance to supervised training with $10^3$ and $10^4$ labels, which are around $50\%$ and $60\%$ reduction in terms of annotation cost, respectively. We also observe that improvement margins diminish as the number of available labels grow significantly (*e.g.* $+0.44dB$ with $10,000$ labels).

**Effect of Residual Contrastive Loss** Following the same training procedure of N2N and N2S, we minimize $\mathcal{L}_{\text{consistency}}$ (Eq. 5) alone to validate the contribution of the residual contrastive loss. The result of minimizing $\mathcal{L}_{\text{consistency}}$ alone is lower than N2N and N2S but higher than VADeR+, DenseCL+, and PixContrast+, which are negatively impacted by the task misalignment effect. It is worth mentioning that including the residual contrastive loss in the training can not significantly improve the model robustness or generalization ability for different downstream tasks, as shown in Table 2.

**Learning from Residuals** It is important to validate that RCL indeed learns from the residuals in the proposed formulation. To illustrate the learning outcome directly, we extract the residual tensors by using a U-Net trained on the MIT dataset with RCL-EMD. Given an anchor image, we calculate the pair-wise difference for EMD between a negative pair and EMD between a positive pair. Given the same network, we record the differences before the training starts (*i.e.* the weights are randomly initialized) and after the loss converges. The density plot of the differences is shown in Fig. 2. RCL contracts the predicted distribution closer to the true underlying distribution, where we use the sampled noise as the residual. We also find that employing large $\alpha$ values in Eq. 6 degrade the performance. We conjecture that this is because low-level vision tasks are sensitive to pixel-level perturbation. To provide an example: a minor change in predicted pixel intensity can change the reconstructed pixel color but an analogous change in predicted pixel probability may not meaningfully change *e.g.* a segmentation result. RCL with large $\alpha$ can still learn representations, however these may not be appropriate for the downstream tasks, discussed in Sec. 3.5.

**Limitations** The empirical results in Sec. 5 are based on simulation. It is interesting to evaluate the proposed framework on real-world multi-task datasets in the future.

## 6 Conclusion

We present a principled unsupervised strategy which can learn transferable representations from images with additive noise for different image reconstruction tasks. To the best of our knowledge, we are the first to unify CL and residual learning by formulating a residual-based instance discrimination pretext task. The empirical studies validate the robustness and generalization of the representations learned by RCL, and further pose a new generic and label-efficient learning direction for low-level vision tasks.

## References


[Batson and Royer, 2019] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *ICML*, pages 524–533. PMLR, 2019.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[Chuang *et al.*, 2020] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NIPS*, volume 33, pages 8765–8775, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[Dong *et al.*, 2021] Nanqing Dong, Michael Kampffmeyer, and Irina Voiculescu. Self-supervised multi-task representation learning for sequential medical images. In *ECML*, pages 779–794. Springer, 2021.

[Ehret *et al.*, 2019] Thibaud Ehret, Axel Davy, Pablo Arias, and Gabriele Facciolo. Joint demosaicking and denoising by fine-tuning of bursts of raw images. In *ICCV*, pages 8868–8877, 2019.

[Ericsson *et al.*, 2021] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, pages 5414–5423, 2021.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[Gharbi *et al.*, 2016] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM TOG*, 35(6):1–12, 2016.

[Hasinoff *et al.*, 2010] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *CVPR*, pages 553–560. IEEE, 2010.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Lehtinen *et al.*, 2018] Jaakko Lehtinen, Jacob Munkberg, and et al. Noise2noise: Learning image restoration without clean data. In *ICML*, pages 2965–2974. PMLR, 2018.

[Li *et al.*, 2018] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018.

[Li *et al.*, 2021] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

[Liu *et al.*, 2020] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020.

[O. Pinheiro *et al.*, 2020] Pedro O O. Pinheiro, Amjad Almahairi, and et al. Unsupervised learning of dense visual representations. In *NIPS*, volume 33, pages 4489–4500. Curran Associates, Inc., 2020.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Wang and Isola, 2020] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939. PMLR, 2020.

[Wang *et al.*, 2021a] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021.

[Wang *et al.*, 2021b] Yanbo Wang, Shaohui Lin, and et al. Towards compact single image super-resolution via contrastive self-distillation. In *IJCAI*, pages 1122–1128, 2021.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[Wu *et al.*, 2021] Haiyan Wu, Yanyun Qu, and et al. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021.

[Xie *et al.*, 2021] Zhenda Xie, Yutong Lin, and et al. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, pages 16684–16693, 2021.

[Zamir *et al.*, 2018] Amir R Zamir, Alexander Sax, and et al. Taskonomy: Disentangling task transfer learning. In *CVPR*, pages 3712–3722, 2018.

[Zhang *et al.*, 2016] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

[Zhang *et al.*, 2017] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017.