International Journal of Automation and Computing
www.ijac.net
DOI: 10.1007/s11633-021-1297-9

Contrastive Self-supervised Representation Learning Using Synthetic Data

Dong-Yu She Kun Xu

Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract: Learning discriminative representations with deep neural networks often relies on massive labeled data, which is expensive and difficult to obtain in many real scenarios. As an alternative, self-supervised learning that leverages input itself as supervision is strongly preferred for its soaring performance on visual representation learning. This paper introduces a contrastive self-supervised framework for learning generalizable representations on the synthetic data that can be obtained easily with complete controllability. Specifically, we propose to optimize a contrastive learning task and a physical property prediction task simultaneously. Given the synthetic scene, the first task aims to maximize agreement between a pair of synthetic images generated by our proposed view sampling module, while the second task aims to predict three physical property maps, i.e., depth, instance contour maps, and surface normal maps. In addition, a feature-level domain adaptation technique with adversarial training is applied to reduce the domain difference between the realistic and the synthetic data. Experiments demonstrate that our proposed method achieves state-of-the-art performance on several visual recognition datasets.

Keywords: Self-supervised learning, contrastive learning, synthetic image, convolutional neural network, representation learning.

Citation: D. Y. She, K. Xu. Contrastive self-supervised representation learning using synthetic data. *International Journal of Automation and Computing*. http://doi.org/10.1007/s11633-021-1297-9

1 Introduction

Convolutional neural networks (ConvNets) have made tremendous progress in the computer vision field[1–3]. However, such achievements are mainly backed up by supervised learning of networks on a massive collection of training data. More recently, various methods[4, 5] try to learn visual representations from large-scale unlabeled data without using any human annotation. A natural solution is self-supervised learning (SSL), which defines an annotation-free surrogate task and uses input itself as the supervision signal[4, 6]. The intuition is that solving tasks like inferring geometrical configuration[7] and recovering missing parts of images[8] can force the ConvNets to learn the semantic representations.

Unlike the existing self-supervised learning that learns representation from realistic data, this paper aims to learn general-purpose visual representations leveraging the synthetic data and their various 'free' annotations. Compared with collecting and annotating photos from the real-world, synthesized data can be easier and cheaper to obtain. For example, it is labor-consuming and impractical to take photos of some objects like birds, while it is feasible to generate a panoramic view of synthetic data. The attributes (e.g., lighting, physics, position) of the synthetic objects can be fully controlled and easily obtained, which can greatly enhance model robustness.

In this work, we present a multi-task self-supervised framework for learning general-purpose visual representations leveraging the semantic information from synthetic data. Specifically, given the synthetic scene, our proposed framework maximizes the agreement between different views of the same scene via a contrastive loss and predicts the free physical cues, including depth, instance contour maps, and surface normal, simultaneously. Besides, to tackle the domain difference between synthetic images and realistic images, we also employ a feature-level domain adaptation technique with adversarial training. Experiments demonstrate that our proposed method achieves state-of-the-art results in self-supervised learning, verifying the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 summarizes the related work on self-supervised learning methods. Section 3 introduces our proposed framework for representation learning on synthetic data. In Section 4, we present the experimental results on the popular benchmark datasets. Finally, Section 5 concludes this paper.

Research Article
Manuscript received October 15, 2020; accepted March 29, 2021
Recommended by Associate Editor Jangmyung Lee
Colored figures are available in the online version at https://link.springer.com/journal/11633

2 Related work

This paper concerns unsupervised representation learning, which learns a representation function mapping the input data to feature vectors without the requirement for semantic labels[9–11]. As a branch of unsupervised learning methods, self-supervised learning refers to learning methods in which deep models are explicitly trained using various cues and proxy tasks[12]. According to the objectives, the existing empirical methods can be categorized into two main types, i.e., predictive methods[5, 8, 13, 14] and contrastive learning methods[15–18].

2.1 Predictive SSL methods

A considerable number of context-based methods have been proposed in recent years, which rely on different context signals, e.g., context similarity, geometric transformation, in either spatial space[5, 19, 20] or temporal space[21, 22]. Fig. 1(a) shows a general pipeline of predictive self-supervised learning. For the context similarity task, the key idea is applying the skip-gram model[23] to the visual domain. Each arrangement is assigned with a class label, and the network aims to solve a supervised problem by predicting the correct arrangement of data patches. For example, Doersch et al.[5] use the relative position of two patches in a set of eight possible spatial configurations, while Noroozi et al.[20, 22] make an extension using 3×3 patches in a Jigsaw puzzle configuration. Besides, Lee et al. [21] use the temporal ordering of patches by shuffling four consecutive video frames to 12 classes for prediction. Zhan et al.[24] propose to embed pixels so that the similarity between the embeddings matches the similarity between their optical flow vectors. In terms of geometric transformation, Gidaris et al.[7] train ConvNets to identify rotations applied to the input image. The basic premise is that predicting rotation teaches neural networks to recognize and localize salient object parts in the image. Feng et al.[25] further decouple predicting rotations from discriminating individual instances.

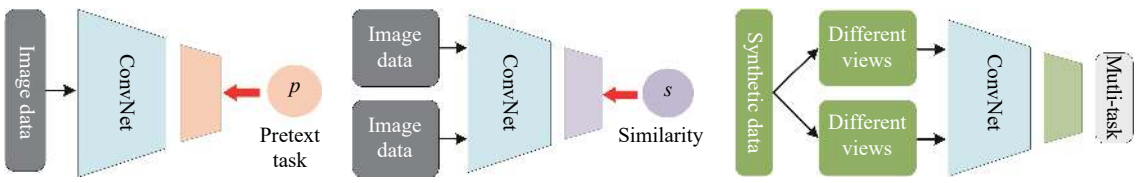
Apart from using a single task, there are several methods considering multiple supervisory signals for representation learning[6, 26–28]. For example, Wang et al.[26] unify different types of in-variance by training two tasks in sequential order. Zhang et al.[29] propose a network with two groups for a bidirectional cross-channel prediction that can aggregate complementary image representations. Doersch and Zisserman[6] consider using multiple self-supervised tasks to obtain a performance boost, while Zhang et al.[27] present the automatic code transformation for unsupervised representation learning. More recently, as synthetic data shows great potential in various vision tasks[30], Ren and Lee[13] use such data to learn self-supervised representations for general vision tasks. However, they only utilize the physical property maps while ignoring the potential semantic information for the synthetic images.

2.2 Contrastive SSL methods

Contrastive methods have led to great empirical success in visual tasks with self-supervised contrastive pre-training[17, 31, 32]. Different from the predictive SSL methods requiring pre-defined tasks, contrastive SSL methods learn representation by contrasting information between positive and negative examples, as shown in Fig. 1(b). Deep InfoMax[17] is the first work using a contrastive learning task to explicitly model mutual information, which aims to maximize the mutual information (MI) between a local patch and its global context. Contrastive predictive coding (CPC)[16] is proposed to maximize the association between a segment of audio and its context audio for speech recognition.

Recently, Tschannen et al.[33] prove that an upper bound MI estimator leads to ill-conditioned representation, pointing that the success of such above methods is more attributed to encoder architecture and metric learning. It is also empirically supported in recent methods leveraging instance discrimination as a pretext task[34, 35]. Contrastive multiview coding (CMC)[36] adopts multiple different views (e.g., luminance, chrominance, depth, and optical flow) of an image as positive samples, while momentum contrast (MoCo)[34] further develop the idea via momentum updating the negative encoder. Furthermore, SimCLR[35] introduces several different forms of data augmentation (e.g., crop and resize, rotate, Gaussian blur) as different views for contrastive learning, which improves considerably over previous methods for self-supervised learning.

In this paper, we follow this research direction and



(a) Predictive SSL model (b) Contrastive SSL model (c) Our proposed model

Fig. 1 Illustration of different learning models: (a) Predictive self-supervised model; (b) Contrastive self-supervised model; (c) Our proposed model. In (a), the pretext tasks need to be specified manually. In (b), a set of paired data from the realistic domain only requires similarity annotation. The proposed model in (c) introduces cross-domain self-supervised learning by incorporating predictive and contrastive learning tasks for synthetic data.

analyze whether such contrastive self-supervised learning can learn effective representation from the synthetic domain. As shown in Fig. 1(c), we propose a self-supervised learning framework to incorporate both contrastive and predictive tasks for learning general-purpose visual representations.

3 Methodology

We propose a self-supervised deep framework to learn generalizable visual representations from the synthetic data. As illustrated in Fig. 2, there are three main components: 1) view sampling module that samples the inputs with correlated views from the same synthetic scene; 2) multi-task self-supervised learning module that includes physical property prediction and contrastive learning for the synthetic data; 3) feature-level domain adaptation technique that minimizes the feature space gap between the realistic and synthetic domain with adversarial training.

Formally, given a synthetic dataset $\mathcal{X}^S = \{\boldsymbol{X}_i\}_{i=1}^{N_s}$ of N_s scenes, we first generate paired input images $V(\boldsymbol{X}_i) = \{(x_{ij}^1, x_{ij}^2)\}_{j=1}^{n}$ for each scene. Here, $V(\cdot)$ and n denote the view sampling module and the number of images sampled from one synthetic scene. Our goal is to learn an encoder network $E : x \mapsto f_x$, which aims to map the input image x to the general-purpose visual representation $E(x; \theta_E)$. Note that θ_E denotes the shared weights of the encoder network, which can be further transferred to the downstream tasks for realistic dataset $\mathcal{Y}^T = \{y_i\}_{i=1}^{N_t}$ of N_t images. The following subsections provide a detailed description of the multi-task SSL module, which consists of contrastive learning and physical property prediction tasks.

3.1 Contrastive learning for multi-view data

3.1.1 Contrastive learning loss

Inspired by recent contrastive learning methods[16, 35, 36], the first proposed task learns visual representation by maximizing agreement between examples with two correlated views using a contrastive learning loss in the latent space. Given N synthetic scenes $\boldsymbol{X_s}$, the paired input images $(x_s^1, x_s^2) = V(\boldsymbol{X_s})$ are first generated by the proposed view sampling module (described in the following subsection). We encode the input to feature representation $f_j = E(x_j)$, where $f_j \in \mathbf{R}^{d_1}$ is the output after the global pooling layer. Then, the features are further mapped to $h_j = G(f_j) \in \mathbf{R}^{d_2}$. After sampling $2N$ synthetic images, we define the contrastive prediction task on the pair-sampled examples from the same minibatch. Following [35, 37], for each positive sample, we treat $2(N-1)$ samples from other synthetic scenes in the minibatch as negative samples rather than sampling negative examples explicitly, which may prevent the network from stacking on the local minimum. Therefore, the contrastive loss function[35, 36] on a positive pair of examples (h_p, h_q) is defined as

$$\ell(p,q) = -\log \frac{\exp\left(\mathrm{sim}\left(h_p, h_q\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\mathrm{sim}\left(h_p, h_k\right)/\tau\right)} \tag{1}$$

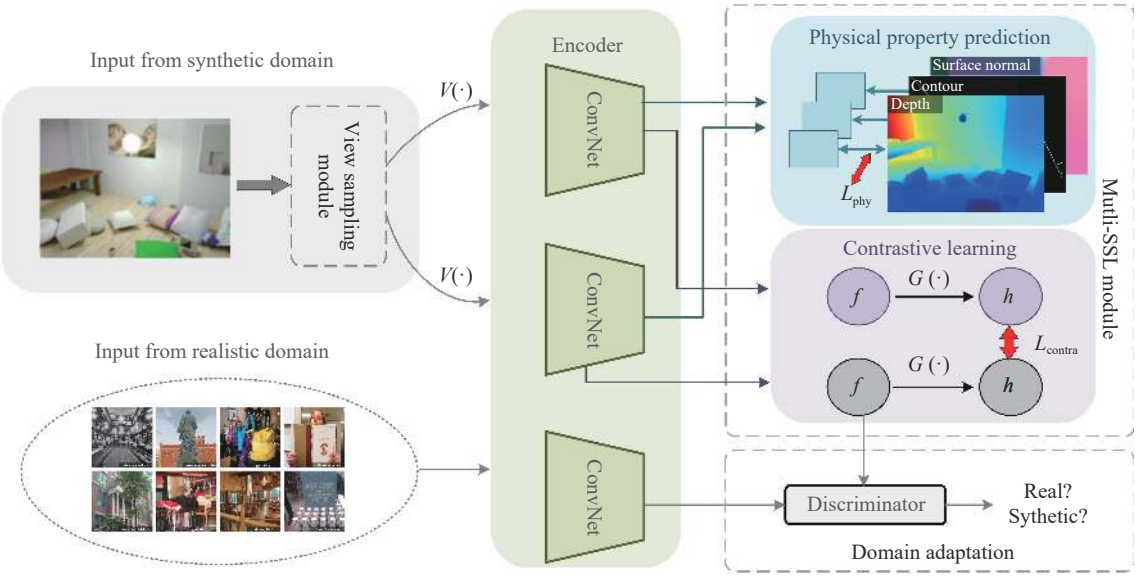


Fig. 2 Our proposed framework for learning self-supervised representation on the synthetic data. Given the synthetic scene, our network first generates paired inputs via the view sampling module. The encoder with shared weights is optimized with the multi-SSL module, including physical property prediction and contrastive learning tasks. For the former task, the ConvNet computes the physical property loss according to its corresponding physical cue, i.e., depth, instance contour map, and surface normal. For the latter task, the ConvNet aims to maximize agreement between paired inputs with two correlated views via the contrastive loss. Meanwhile, we use the discriminator to differentiate features from the realistic and synthetic images, aiming to minimize the feature space gap. The learned green ConvNet can be further used for transfer learning on the downstream tasks for real images.

where $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^{\top}\boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denote the cosine similarity between two vectors extracted from the same scene, and $\mathbf{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function. Then, the final contrastive loss is computed across all positive pairs defined as

$$L_{contra} = \frac{1}{2N}\sum_{k=1}^{N}[\ell(2k-1, 2k) + \ell(2k, 2k-1)]. \quad (2)$$

3.1.2 View sampling module

Previous methods[6, 35] employ various transformations on the realistic data for data augmentation (e.g., crop, rotation, color distortion) to utilize supervision from the raw data. Compared with realistic images captured from one viewpoint, synthetic data have more degrees of freedom and thus more flexible ways of transformation, especially for the viewpoint.

In this work, we propose the view sampling module to generate different views in the synthetic scene following the same routine in the SceneNet RGBD dataset[38]. First, given a scene type, the objects are selected according to the distribution of object categories of such type in the SUN-RGBD real-world dataset. Then an off-the-shelf physics engine, Project Chrono[1], is employed to simulate the scene. Second, the simple random trajectory paths are generated by simulating two physical bodies. The first body decides the location of the camera (i.e., viewpoint) in the scene, while the second body acts as a proxy for a human paying attention to random points. Given the input scene $\boldsymbol{X}_i$, the random trajectories can be generated by simulating motions of these bodies with Euler integration and applying 3D directional force vectors randomly. For each sampling image x_{ij}^1 corresponding with a position and the pose, we sample an additional image x_{ij}^2 by applying limited rotational freedom $V(\cdot)$ on the pose while fixing the position. Specifically, we sample the different views by changing yaw and pitch slightly and locking roll entirely. Finally, the opposite renderer[39] is employed to produce photorealistic rendering using photon mapping. Similarly, we assign textures randomly to each of its constituent components and add random lighting on the scene to improve the variability.

3.2 Physical property prediction

To leverage the free knowledge from the synthetic domain, the proposed framework also aims to predict three corresponding physical cues of synthetic input images following Ren and Lee[13]. Predicting such physical maps requires the ConvNet to understand high-level semantics about the relative placements and contours of the objects in a scene, such a task offers a promising supervisory signal for self-supervised learning. First, for the depth prediction, we explicitly add a convolutional layer to predict the depth maps $\hat{M}^d \in \mathbf{R}_{w\times h\times 1}$. Assume the ground truth depth maps are $M^d \in \mathbf{R}_{w\times h\times 1}$, so the depth prediction loss can be computed by

$$L_{depth} = \frac{1}{m}\sum_i {t_i}^2 - \frac{1}{2m^2}\sum_{i,j} t_i t_j \quad (3)$$

where $m = w \times h$ and $t = \log M^d - \log \hat{M}^d$ is the element-wise difference between log depth maps.

Second, for the instance contour prediction, the ground truth maps are detected using a canny edge detector for instance-level segmentation. Thus, this sub-task is formulated as a binary semantic edge/non-edge prediction task. We note the predicted instance contour map and ground truth map as $\hat{M}^c, M^c \in \mathbf{R}_{w\times h\times 1}$. Thus, the class-balanced sigmoid cross entropy loss[40] is defined as

$$L_{contour} = -\beta\sum_{i\in M_+^c}\log Pr\left(\hat{M}_i^c = 1\right) - (1-\beta)\sum_{i\in M_-^c}\log Pr\left(\hat{M}_i^c = 0\right) \quad (4)$$

where M_-^c and M_+^c denote the number of ground-truth edges and the number of non-edges sets, $\beta = |M_-^c| / |M_-^c + M_+^c|$. Here, $Pr\left(\hat{M}^c{}_i = 1\right) \in [0, 1]$ is computed using the sigmoid function on the activation value, which means the probabilities for a predicted pixel belong to an edge.

Third, for the surface normal, we also use a convolutional layer to predict the surface normal map $\hat{M}^n \in \mathbf{R}_{w\times h\times 3}$. Given the ground truth M^n, the surface normal prediction loss is computed via dot product as follows:

$$L_{normal} = -\frac{1}{n}\sum_i \hat{M}_i^n \cdot M_i^n. \quad (5)$$

Therefore, the final physical property task is formulated with a weighted multi-task loss, defined as

$$L_{phy} = \alpha_d L_{depth} + \alpha_c L_{contour} + \alpha_n L_{normal} \quad (6)$$

where the weights are used to scale these terms with similar magnitude.

3.3 Domain adaptation with adversarial loss

We assume that there exist two distributions derived from the synthetic dataset $\mathcal{X}^S = \{\boldsymbol{X}_i\}_{i=1}^{N_s}$ of N_s scenes and the realistic dataset $\mathcal{Y}^T = \{y_i\}_{i=1}^{N_t}$ of N_t images, referred to the source domain and target domain, respectively. Obviously, $\mathcal{X}^S$ is shifted from $\mathcal{Y}^T$ by some domain shift. To address the problem of domain difference, we employ the general adversarial unsupervised adaptation

[1] https://projectchrono.org/

methods[41–45] to reduce the gap between real and synthetic data. For the training images from both the source and the target domains, we first obtain the feature vectors with a weight-shared encoder network. Then, we train a domain discriminator, denoted with D, to perform an unsupervised domain adaptation from synthetic to real data based on adversarial learning. We assign images from the source distribution and target distribution with label 1 and label 0, respectively. Thus, the parameters of the discriminator θ_D can be optimized by minimizing the following binary cross-entropy loss:

$$L_D\left(\theta_D \mid f_{\boldsymbol{x}}, f_{\boldsymbol{y}}\right) = -\sum_i \log\left(D\left(f_{x_i}\right)\right) - \sum_j \log\left(1 - D\left(f_{y_j}\right)\right). \tag{7}$$

3.4 Optimization

Overall, the discriminator network E maps both real and synthetic data to the feature as visual representation, while the domain discriminator D tries to differentiate real and synthetic features. The whole framework is optimized by updating the encoder network E and discriminator network D alternatively.

In specific, we first keep D unchanged and update E as well as the mapping layers by minimizing the following loss function:

$$L_E\left(\theta_E \mid f_{\mathbf{x}}\right) = -\sum_i \log\left(1 - D\left(f_{x_i}\right)\right) + \lambda L_{phy} + (1-\lambda) L_{contra} \tag{8}$$

where λ is the trade-off between two SSL tasks.

For the next stage, we fix E and optimize D with (7). By repeatedly doing so, E and D can learn from each other to minimize the domain difference between synthetic and real-world images so that the features learned from synthetic images can generalize to real images. Since all the parameters can be derived, we are able to present an effective representation learning using stochastic gradient descent (SGD) for network optimization.

4 Experiments

In this section, we evaluate the proposed method on several transfer learning benchmarks, including fine-tuning for PASCAL visual obiect classes challenge (PASCAL VOC) classification and detection, linear and nonlinear classification on ImageNet and Places, and surface normal estimation on NYUv2 RGBD dataset (NYUD). We first report the results on several standard transfer learning benchmarks. Then, we perform ablation studies to show the effect of the individual components of the proposed method. Our experiments illustrate that the pre-training model on the multi-task self-supervised tasks yields features that outperform the previous self-supervised learning methods.

4.1 Implementation details

Architecture. Our framework is based on the AlexNet architecture[46] for a fair comparison with the previous methods[5, 20]. The input images are first resized to the size of 256×256 and then randomly cropped to 227×227. We also duplicate one of the RGB channels three times, resulting in grayscale inputs for learning more robust features following[5, 13, 26]. The hyperparameter λ representing a tradeoff between two SSL tasks is set to 0.7 in the following experiments, which will be discussed in Section 4.3. In addition, for α_d, α_c and α_n in (6), we set such hyperparameters as 5, 1 and 10, respectively, mainly for scaling the gradients to have similar magnitude. Following [35], the temperature parameter τ in (1) is set to 0.1, and the feature dimensions d_1 and d_2 are set to 4 096 and 2 048, respectively. The proposed network is first initialized randomly and trained for 50 epochs with an initial learning rate of 0.01. We use a weight decay of 0.000 5 and optimize the whole network with SGD. For the supervised learning task, we initialize the AlexNet model with the weights from the convolutional layers learned by the self-supervised model.

Dataset. We use SceneNet RGB-D[38] dataset as the synthetic images and generate the images from different views with a view sampling module. The ground truth depth maps and surface normal maps are provided for each synthetic image by [4, 38], respectively. The instance contour maps can be computed using a canny edge detector following [13]. In addition, we use images from Places365-Standard[47] as the realistic data, which includes 1.8 million images comprising more than 400 unique scene categories for training. Dataset examples are shown in Fig. 3.

Baselines. We enumerate several alternative pretext tasks that use images and their self-supervisions for representation learning.

1) Pathak et al.[8] (Inpainting) learn to recover missing pixels of images with a generative adversarial model.

2) Zhang et al.[14] (Colorization) learn representation by mapping from a grayscale input to a distribution over quantized color value outputs with multinomial cross-entropy loss.

3) Doersch et al.[5] (Position) train the eight-class classification ConvNet to predict the location of two patches sampled from the input images, which feeds two input patches and fuses the output to assign a probability to each of the eight spatial configurations.

4) Noroozi and Favaro[20] (Jigsaw puzzles) train the ConvNet to solve the 3×3 Jigsaw puzzles with a predefined permutation set, which is further boosted by incorporating occlusions in the tiles in Jigsaw++[48].

5) Zhang et al.[29] (Split-brain) train the split-brain

(a) SceneNet-RGBD

(b) Places365

Fig. 3 Example images from (a) synthetic domain and (b) realistic domain. Both datasets contain images covering a variety of indoor scenarios, while Places365 also includes outdoor scenes.

autoencoder network with one half performing colorization and the other half performing grayscale prediction with a cross-entropy objective.

6) Gidaris et al.[7] (RotNet) train the ConvNet to recognize the four possible geometric transformations that are applied to the image, i.e., the 0, 90, 180 and 270 degrees rotations.

Note that all these compared SLL methods are trained using realistic datasets from scratch, while cross-domain[13] and ours are pre-trained using the synthetic dataset and evaluated on the datasets from the realistic domain, i.e., ImageNet, PASCAL VOC and Places datasets.

4.2 Transfer learning evaluation

4.2.1 Classification on ImageNet and Places

We evaluate the generalization of our self-supervised learned features by training with both linear and nonlinear object classifiers for the ImageNet classification task. First, we train linear classifiers on top of the features extracted by different convolutional layers, where the layers transferred from self-supervised trained models are fixed during training. Such linear classification can directly evaluate the discriminative power of the learned representation over the object class. We illustrate the performance on ImageNet and Places in Tables 1 and 2, respectively. As observed, our proposed framework is comparable to self-supervised methods, which shows improvement over models initialized randomly and using data-dependent initialization[49] The reason why our method underperforms the RotNet method is that most SSL methods are trained directly using images from the ImageNet dataset. In contrast, our method is trained using synthetic indoor images without seeing any object from the ImageNet. Note that [13] is the only cross-domain baseline, while our method outperforms this method by a large margin, about 4.5% on conv4. The results illustrate the effectiveness of the proposed contrastive learning framework for the synthetic data. For the Places dataset with more similar high-level semantics, our proposed method outperforms most self-supervised methods, showing that the learned representation is helpful for object recognition in the realistic domain.

Second, we perform nonlinear classification on ImageNet by freezing several layers and training the remaining layers from scratch. This experiment can illustrate the alignment between the discrimination ability and the ground truth class. As lower layers mainly capture low-level information (e.g., contours and edges), these features are with relatively low accuracies and generally less often used. We report the results of different self-supervised approaches freezing layers from the first layer to conv4 or conv5. As shown in Table 3, our method

Table 1 ImageNet top-1 classification with a linear classifier. We compare our self-supervised feature learning approach with other approaches by training logistic regression classifiers on top of the feature maps of each layer to perform object recognition tasks. The ImageNet-labels AlexNet and Gaussian-init AlexNet are initialized with parameters pre-trained on the ImageNet dataset and initialized randomly, respectively. Note that all compared methods use AlexNet as the backbone, and * represents training self-supervised model without using ImageNet images.

Method	Conv1	Conv2	Conv3	Conv4	Conv5
ImageNet-labels AlexNet[46]	19.3	36.3	44.2	48.3	50.5
Gaussian-init AlexNet[46]	11.6	17.1	16.9	16.3	14.1
Krähenbühl et al.[49]	17.5	23.0	24.5	23.2	20.6
Pathak et al. (Inpainting)[8]	14.1	20.7	21.0	19.8	15.5
Zhang et al. (Colorization)[14]	12.5	24.5	30.4	31.5	30.3
Doersch et al. (Position)[5]	16.2	23.3	30.2	31.7	29.6
Noroozi and Favaro (Jigsaw puzzles)[20]	18.2	28.8	34.0	33.9	27.1
Noroozi et al. (Jigsaw++)[48]	18.2	28.7	34.1	33.2	28.0
Noroozi et al. (Counting)[50]	18.0	30.6	34.3	32.5	25.7
Zhang et al. (Split-brain)[29]	17.7	29.3	35.4	35.2	32.8
Gidaris et al. (RotNet)[7]	18.8	31.7	38.7	38.2	36.5
Ren and Lee (Cross-domain)*[13]	16.5	27.0	30.5	30.1	26.5
Ours*	17.2	30.2	34.2	35.6	31.5

Table 2 Places classification with a linear classifier. We compare our self-supervised feature learning approach with other approaches by training logistic regression classifiers on top of the feature maps of each layer to perform object recognition tasks. The ImageNet-labels AlexNet and Gaussian-init AlexNet are initialized with parameters pre-trained on the ImageNet dataset and initialized randomly, respectively. Note that all compared methods use AlexNet as the backbone.

Method	Conv1	Conv2	Conv3	Conv4	Conv5
Places-labels AlexNet[51]	22.1	35.1	40.2	43.3	44.6
Gaussian-init AlexNet[46]	15.7	20.3	19.8	19.1	17.5
ImageNet-labels AlexNet[46]	22.7	34.8	38.4	39.4	38.7
Zhang et al. (Colorization)[14]	16.0	25.7	29.6	30.3	29.7
Doersch et al. (Position)[5]	19.7	26.7	31.9	32.7	30.9
Noroozi and Favaro (Jigsaw puzzles)[20]	23.0	31.9	35.0	34.2	29.3
Noroozi et al. (Jigsaw++)[48]	22.0	31.2	34.3	33.9	22.9
Noroozi et al. (Counting)[50]	23.3	33.9	36.3	34.7	29.6
Zhang et al. (Split-brain)[29]	21.3	30.7	34.0	34.1	32.5
Gidaris et al. (RotNet)[7]	18.8	31.7	38.7	38.2	36.5
Ours	21.6	32.7	37.5	36.1	34.5

achieves comparable results on conv4 and conv5 layers with the baseline methods, illustrating that the proposed self-supervised task is able to learn a discriminative representation.

4.2.2 Classification, object detection and semantic segmentation on PASCAL VOC

We evaluate the transferability of the learned representation on the PASCAL VOC dataset[52]. Similar to Section 4.2.1, we initialize the model with our self-supervised learning model and then fine-tune the model for PASCAL VOC classification, detection, and segmentation tasks. Classification and detection tasks are measured by mean average precision (mAP), and segmentation task is measured by mean intersection over union (mIoU). The results are reported in Table 4. We fine-tune the model on PASCAL VOC 2007 and perform multi-label classification on the test set for the classification task. Our method can improve upon [13], while underperforming the state-of-the-art SSL method mainly due to that our self-training does not utilize images from PASCAL VOC. For the detection task, we fine-tune Fast-RCNN[53] initialized with our self trained network using multi-scale training and single-scale testing. We further improve the mAP by 1.3% compared with [13]. In addition, we also fine-tune our model with fully convolutional network

Table 3 ImageNet top-1 classification results with non-linear classifiers using different self-supervised feature learning methods. The ImageNet AlexNet and Random-init AlexNet are pre-trained on the ImageNet dataset and initialized randomly, respectively, while the other networks are pre-trained with different self-supervisions. Note that * represents that training self-supervised model without using ImageNet images.

Method	Conv4	Conv5
ImageNet-labels AlexNet[46]	59.7	59.7
Gaussian-init AlexNet[46]	27.1	12.0
Zhang et al. (Colorization)[14]	40.7	35.2
Doersch et al. (Position)[5]	45.6	30.4
Noroozi and Favaro (Jigsaw Puzzles)[20]	45.3	34.6
Noroozi et al. (Jigsaw++)[49]	46.1	35.4
Noroozi et al. (Counting)[50]	43.3	32.9
Gidaris et al. (RotNet)[7]	50.0	43.8
Ours*	46.5	33.9

Table 4 Classification, detection, and segmentation results on the PASCAL VOC dataset. The unsupervised methods are pre-trained on the ImageNet without using the semantic labels. We report the mean average precision on the classification and detection tasks and the mean intersection over union on the segmentation task. Note that we fine-tune the whole model for all three tasks and use multi-scale training and single-scale testing for detection. Here, † represents training self-supervised model without using PASCAL VOC images.

Method	Classification	Detection	Segmentation
ImageNet-labels AlexNet[46]	79.9	56.8	48.0
Gaussian-init AlexNet[46]	53.3	43.4	19.8
Krähenbühl et al.[48]	56.5	45.6	32.6
Pathak et al. (Inpainting) [8]	56.4	44.5	29.7
Zhang et al. (Colorization)[14]	65.6	46.9	35.6
Doersch et al. (Position)[5]	65.3	51.1	–
Noroozi and Favaro (Jigsaw Puzzles)[20]	67.6	53.2	37.6
Noroozi et al. (Counting)[50]	67.7	51.4	36.6
Zhang et al. (Split-Brain)[29]	67.1	46.7	36.0
Gidaris et al. (RotNet)[7]	72.9	54.4	39.1
Ren and Lee (Cross-domain)† [13]	68.0	52.6	–
Ours†	69.2	53.9	39.0

(FCN)[54] architecture for the segmentation task. It is important to note that our method achieves state-of-the-art on segmentation, which illustrates that our learned self-supervised representation can be generalized across tasks with different abstract semantic.

4.2.3 Surface normal estimation on NYUD

To study how well our self-supervised framework learns representation for the synthetic scene, we further evaluate the surface normal estimation performance on the NYUv2 RGBD dataset[55]. Following the protocol in [13], we report the transfer learning results by first pre-training on the proposed multi-SSL tasks and then fine-tuning for surface normal estimation. We use the standard split of 795 images for training and 654 images for testing.

We compare with several baselines and self-supervised models[13, 26]. Wang et al.[26] apply the FCN architecture followed with a codebook of 40 codewords to encode the 3-dimension normals, which is pre-trained on pairs of images exhibiting richer visual invariance. Following [13], our model is pre-trained on 0.5 million SUNCG synthetic images using the FCN architecture with skip connections. The model trained from scratch and ImageNet pre-trained model also use similar FCN[26], which are initialized with Xavier initialization and parameters pre-trained on ImageNet, respectively. All the compared models are then fine-tuned on the NYU dataset using the ground-truth surface normal provided in [56]. The results are shown in Table 5. It can be seen that our proposed self-supervised model achieves the best performance illustrating the effectiveness of learned representation for the synthetic task. Surprisingly, compared with the ImageNet pre-trained model, our model improves the percentage of pixels with an error less than 30° from 63.4 to 75.3 sig-

nificantly.

4.3 Ablation study

We conduct an ablation study on PASCAL VOC classification to see the influence of each component (i.e., view sampling module, multi-task SSL, domain adaptation) in the proposed framework.

Multiple SSL tasks. We first evaluate the performance of only using one single SSL task for representation learning. As reported in the first two rows of Table 6, the representation learned with the physical property prediction task is sub-optimal compared with the contrastive learning task. We also report the results of combining both tasks with different λ in the middle five rows. When λ increases from 0 to 0.7, the classification performance is boosted dramatically since these SSL tasks are complementary for representation learning. As further increasing λ leads to decreased accuracy, we set $\lambda = 0.7$ in all our experiments.

View sampling module. We also evaluate the effect of the proposed sampling module for the synthetic data. We employ a simple random sampling strategy as a baseline, which samples two frames in a random camera trajectory[38]. The eighth row in Table 6 shows the result of replacing our proposed view sampling module, where the accuracy has significantly declined by about 4.4%. The reason is that our proposed view sampling module is more effective in generating inputs with convenient distribution for contrastive learning.

Feature adaptation. Finally, we evaluate the transfer learning performance with and without domain adaptation in the last two rows of Table 6. As it can be observed, the performance of the model without feature adaptation drops by about 1.9% in accuracy, mainly due to the data biases between the synthetic and real-world image datasets. By applying the adversarial training, our framework can learn more generalized representation for the realistic tasks.

5 Conclusions

In this work, we address the problem of self-supervised visual representation learning. In specific, we present a multi-task self-supervised framework for contrastive learning of visual representations leveraging the

Table 5 NYUv2 RGBD surface normal estimation using different self-supervised feature learning methods. We report the mean RMSE error (Mean) and median RMSE error (Median) for all visible pixels (in degrees), RMSE means root mean squared error. These two measurements are the error measurements, and lower is better. We also report the percentage of pixels with errors less than 11.25°, 22.5° and 30°. For these three measurements, higher is better.

Method	Mean	Median	11.25°	22.5°	30°
Scratch	31.3	25.3	24.2	45.6	56.8
ImageNet	27.8	21.2	29.0	52.3	63.4
Wang et al.[26]	26.0	18.0	33.9	57.6	67.5
Ren and Lee[13]	23.8	16.2	36.6	62.0	72.9
Ours	22.9	15.5	37.7	63.1	75.3

Table 6 Ablation study results on PASCAL VOC. We evaluate the framework with different choices on individual components. Here, L_{phy} and L_{contra} represent using different SSL tasks for network optimization; R.S. and V.S. denote sampling paired inputs from the synthetic scene using the random sampling strategy and the proposed view sampling module. And DA represents employing feature adaptation.

λ	L_{phy}	L_{contra}	R.S.	V.S.	DA	Accuracy (%)
1	√			–		65.2
0		√		√		66.4
0	√	√		√	√	67.3
0.3	√	√		√	√	68.5
0.5	√	√		√	√	68.8
0.7	√	√		√	√	69.2
1	√	√		√	√	66.3
0.7	√	√	√		√	64.8
0.7	√	√		√		67.3
0.7	√	√		√	√	69.2

semantic information from synthetic data. The key idea is that by solving such tasks, the models are forced to learn not only how objects are assembled in appearance but also what information is shared among different domains. We also employ a feature-level domain adaptation technique with adversarial training, resulting in general-purpose visual representations that can be transferred to real-world tasks. The experiments demonstrate that our proposed method achieves state-of-the-art results in self-supervised learning.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61822204 and 61521002).

Open Access

References

[1] B. Zhao, J. S. Feng, X. Wu, S. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[2] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. Masero, A. Hussain. Deep learning based single image super-resolution: A survey. *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 413–426, 2019. DOI: 10.1007/s11633-019-1183-x.

[3] K. Aukkapinyo, S. Sawangwong, P. Pooyoi, W. Kusakunniran. Localization and classification of rice-grain images using region proposals-based convolutional neural network. *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 233–246, 2020. DOI: 10.1007/s11633-019-1207-6.

[4] X. L. Wang, A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 2794–2802, 2015. DOI: 10.1109/ICCV.2015.320.

[5] C. Doersch, A. Gupta, A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1422–1430, 2015. DOI: 10.1109/ICCV.2015.167.

[6] C. Doersch, A. Zisserman. Multi-task self-supervised visual learning. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 2070–2079, 2017. DOI: 10.1109/ICCV.2017.226.

[7] S. Gidaris, P. Singh, N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 2536–2544, 2016. DOI: 10.1109/CVPR.2016.278.

[9] G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, vol. 313, no. 5786, pp. 504–507, 2006. DOI: 10.1126/science.1127647.

[10] P. Vincent, H. Larochelle, Y. Bengio, P. A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*, ACM, Helsinki, Finland, pp. 1096–1103, 2008. DOI: 10.1145/1390156.1390294.

[11] R. Lopez, J. Regier, M. I. Jordan, N. Yosef. Information constraints on auto-encoding variational bayes. In *Advances in Neural Information Processing*, Montreal, Canada, pp. 6117–6128, 2018.

[12] X. Liu, F. J. Zhang, Z. Y. Hou, Z. Y. Wang, L. Mian, J. Zhang, J. Tang. Self-supervised learning: Generative or contrastive. [Online], Available: https://arxiv.org/abs/2006.08218, 2020.

[13] Z. Z. Ren, Y. Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, pp. 762–771, 2018. DOI: 10.1109/CVPR.2018.00086.

[14] R. Zhang, P. Isola, A. A. Efros. Colorful image colorization. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 649–666, 2016. DOI: 10.1007/978-3-319-46487-9_40.

[15] R. Hadsell, S. Chopra, Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern*, IEEE, New York, USA, pp. 1735–1742, 2006. DOI: 10.1109/CVPR.2006.100.

[16] A. van den Oord, Y. Z. Li, O. Vinyals. Representation learning with contrastive predictive coding. [Online], Available: https://arxiv.org/abs/1807.03748, 2018.

[17] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[18] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, USA, pp. 5628–5637, 2019.

[19] T. Nathan Mundhenk, D. Ho, B. Y. Chen. Improvements to context based self-supervised learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 9339–9348, 2018. DOI: 10.1109/CVPR.2018.00973.

[20] M. Noroozi, P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 69–84, 2016. DOI: 10.1007/978-3-319-46466-4_5.

[21] H. Y. Lee, J. B. Huang, M. Singh, M. H. Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 667–676, 2017. DOI: 10.1109/ICCV.2017.79.

[22] D. Kim, D. Cho, D. Yoo, I. S. Kweon. Learning image representations by completing damaged jigsaw puzzles. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, IEEE, Lake Tahoe, USA, pp. 793–802, 2018. DOI: 10.1109/WACV.2018.00092.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, ACM, Lake Tahoe, USA, pp. 3111–3119, 2013.

[24] X. H. Zhan, X. G. Pan, Z. W. Liu, D. H. Lin, C. C. Loy. Self-supervised learning via conditional motion propagation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 1881–1889, 2019. DOI: 10.1109/CVPR.2019.00198.

[25] Z. Y. Feng, C. Xu, D. C. Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 10364–10374, 2019. DOI: 10.1109/CVPR.2019.01061.

[26] X. L. Wang, K. M. He, A. Gupta. Transitive invariance for self-supervised visual representation learning. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 1338–1347, 2017. DOI: 10.1109/ICCV.2017.149.

[27] L. H. Zhang, G. J. Qi, L. Q. Wang, J. B. Luo. AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 2542–2550, 2019. DOI: 10.1109/CVPR.2019.00265.

[28] J. Donahue, K. Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 10541–10551, 2019.

[29] R. Zhang, P. Isola, A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 645–654, 2017. DOI: 10.1109/CVPR.2017.76.

[30] X. C. Peng, B. C. Sun, K. Ali, K. Saenko. Learning deep object detectors from 3D models. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1278–1286, 2015. DOI: 10.1109/ICCV.2015.151.

[31] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, A. van den Oord. Data-efficient image recognition with contrastive predictive coding. [Online], Available: https://arxiv.org/abs/1905.09272, 2019.

[32] P. Bachman, R. D. Hjelm, W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 15509–15519, 2019.

[33] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, M. Lucic. On mutual information maximization for representation learning. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[34] K. M. He, H. Q. Fan, Y. X. Wu, S. N. Xie, R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9726–9735, 2020. DOI: 10.1109/CVPR42600.2020.00975.

[35] T. Chen, S. Kornblith, M. Norouzi, G. Hinton. A simple framework for contrastive learning of visual representations. [Online], Available: https://arxiv.org/abs/2002.05709, 2020.

[36] Y. L. Tian, D. Krishnan, P. Isola. Contrastive Multiview coding. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 776–794, 2020. DOI: 10.1007/978-3-030-58621-8_45.

[37] T. Chen, Y. Z. Sun, Y. Shi, L. J. Hong. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Halifax, Canada, pp. 767–776, 2017. DOI: 10.1145/3097983.3098202.

[38] J. McCormac, A. Handa, S. Leutenegger, A. J. Davison. SceneNet RGB-D: Can 5M synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 2697–2706, 2017. DOI: 10.1109/ICCV.2017.292.

[39] T. Hachisuka, H. W. Jensen. Parallel progressive photon mapping on GPUS. In *ACM SIGGRAPH ASIA*, Seoul, Proceedings of Korea, pp. 54:1, 2010.

[40] S. N. Xie, Z. W. Tu. Holistically-nested edge detection. *International Journal of Computer Vision*, vol. 125, no. 1–3, pp. 3–18, 2017. DOI: 10.1007/s11263-017-1004-z.

[41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ACM, Montreal, Canada, pp. 2672–2680, 2014.

[42] Y. Ganin, V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 1180–1189, 2015.

[43] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 3722–3731, 2017. DOI: 10.1109/CVPR.2017.18.

[44] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 7167–7176, 2017. DOI: 10.1109/CVPR.2017.316.

[45] K. Sohn, W. L. Shang, X. Yu, M. Chandraker. Unsuper-

vised domain adaptation for distance metric learning. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[46] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ACM, Lake Tahoe, USA, pp. 1097–1105, 2012.

[47] B. L. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018. DOI: 10.1109/TPAMI.2017.2723009.

[48] M. Noroozi, A. Vinjimoor, P. Favaro, H. Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 9359–9367, 2018. DOI: 10.1109/CVPR.2018.00975.

[49] P. Krähenbühl, C. Doersch, J. Donahue, T. Darrell. Data-dependent initializations of convolutional neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

[50] M. Noroozi, H. Pirsiavash, P. Favaro. Representation learning by learning to count. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 5899–5907, 2017. DOI: 10.1109/ICCV.2017.628.

[51] B. Zhou, À. Lapedriza, J. X. Xiao, A. Torralba, A. Oliva. Learning deep features for scene recognition using places database. In *Proceedings of Conference in Neural Information Processing Systems*, Montreal, Canada, pp. 487–495, 2014.

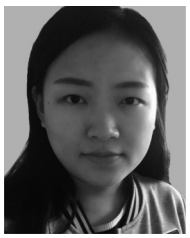
[52] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015. DOI: 10.1007/s11263-014-0733-5.

[53] R. Girshick. Fast R-CNN. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1440–1448, 2015. DOI: 10.1109/ICCV.2015.169.

[54] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 3431–3440, 2015. DOI: 10.1109/CVPR.2015.7298965.

[55] N. Silberman, D. Hoiem, P. Kohli, R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 746-760, 2012. DOI: 10.1007/978-3-642-33715-4_54.

[56] L. Ladicky, B. Zeisl, M. Pollefeys. Discriminatively trained dense surface normal estimation. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 468–484, 2014. DOI: 10.1007/978-3-319-10602-1_31.



Dong-Yu She received the B. Eng. and the M. Eng. degrees in computer science and technology from Nankai University, China in 2019 and 2016, respectively. She is a Ph. D. degree candidate in Department of Computer Science and Technology, Tsinghua University, China.

Her research interests include deep learning and computer vision.

E-mail: shedy19@mails.tsinghua.edu.cn
ORCID iD: 0000-0002-1434-562X

Kun Xu received B. Eng. and Ph.D. degrees in computer science and technology from Tsinghua University, China in 2005 and 2009, respectively. He is an associate professor in Department of Computer Science and Technology, Tsinghua University, China.

His research interests include realistic rendering and image/video editing.

E-mail: xukun@tsinghua.edu.cn (Corresponding author)
ORCID iD: 0000-0002-2671-4170