2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Improving Unimodal Object Recognition with Multimodal Contrastive Learning

Johannes Meyer[1], Andreas Eitel[1], Thomas Brox[1], and Wolfram Burgard[1,2]

***Abstract*— Robots perceive their environment using various sensor modalities, e.g., vision, depth, sound or touch. Each modality provides complementary information for perception. However, while it can be assumed that all modalities are available for training, when deploying the robot in real-world scenarios the sensor setup often varies. In order to gain flexibility with respect to the deployed sensor setup we propose a new multimodal approach within the framework of contrastive learning. In particular, we consider the case of learning from RGB-D images while testing with one modality available, i.e., exclusively RGB or depth. We leverage contrastive learning to capture high-level information between different modalities in a compact feature embedding. We extensively evaluate our multimodal contrastive learning method on the Falling Things dataset and learn representations that outperform prior methods for RGB-D object recognition on the NYU-D dataset. Our code and details on the used datasets are available at: `https://github.com/meyerjo/MultiModalContrastiveLearning`.**

I. Introduction

Object recognition is at the core of many robot applications. For example, a robot needs to know the category of an object in order to grasp a specific object or to adjust its navigation in the vicinity of certain objects. Multimodal object perception has received great attention in recent years [1], [2], [3], [4]. Especially in robotics, we see many applications where multiple sensors and different modalities are used [5], [6], [7]. Multimodal object perception increases robustness because different sensors can provide complementary information.

Nevertheless, multimodal learning comes with a few assumptions that are not addressed thoroughly in prior work. Especially, multimodal learning often assumes that the sensor setup used for training a model will also be available in the exact same configuration during deployment of the robot. However, this assumption does not always hold true. Let's take as an example robots that operate in various warehouses. Ideally, you would equip all robots with exactly the same sensor setup. In practice, the sensor setup might have to vary when moving from one warehouse (or customer) to another. In particular, some sensors are more expensive than others (e.g., LiDAR) and therefore not all robots can have the same sensor setup. In addition, privacy concerns may restrict the usage of sensors in certain environments (e.g., cameras). We conclude that one of the main challenges for multimodal learning is flexibility with respect to a sensor setup that can differ between training and real-world deployment. In this work, we study the case of multimodal learning for unimodal object recognition (i.e., we assume to have access to a multimodal sensor setup during training but only a single sensor during testing). The main question that we ask is: do we benefit from training with multiple modalities (e.g., vision and depth) given that we only have access to one modality at test-time?

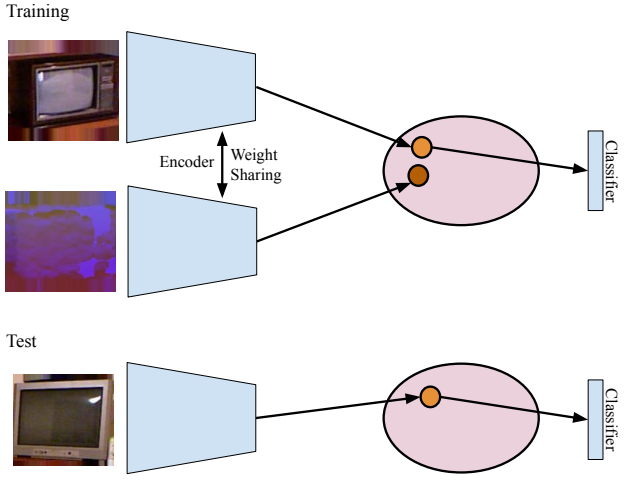

Fig. 1. Our multimodal contrastive learning method maps two images from different modalities into one embedding space using a weight-sharing network. During multimodal training a contrastive loss aims to encode high-level features between both modalities. We propose a multi-task objective that learns to map multimodal images to the embedding space and uses the embedding to classify objects. Following our approach, we can improve recognition performance when using a single modality at test time.

A familiar research question has been addressed in the machine learning community under the paradigm of learning with privileged information. In this learning paradigm "some additional information $\mathbf{x}^\star$ about training example $\mathbf{x}$ [is provided]; this privileged information will not be available at the test stage" [8]. Hence, the key idea is to leverage as much information as possible from the stream of privileged information.

We follow this underlying idea and make several contributions: 1) We present a novel method for learning with privileged information based on a recent technique for self-supervised contrastive learning [9]. We learn multimodal representations using contrastive learning objectives that map same concepts from different modalities to nearby points

This work has been supported by the Freiburg Graduate School of Robotics and by the Federal Ministry of Education and Research (BMBF) under Deep PTL.

[1]All authors are with the Department of Computer Science, University of Freiburg, Germany. meyerjo,eitel,brox,burgard@cs.uni-freiburg.de

[2]Wolfram Burgard is also with the Toyota Research Institute, Los Altos, USA.

in the embedding space, see Figure 1. 2) We propose a novel multi-task objective that combines the multimodal contrastive loss with a unimodal cross-entropy loss in an end-to-end manner. 3) We study the effects of recent data augmentation techniques for both RGB and depth domain on the contrastive learning task. 4) We evaluate our method on publicly available RGB-D datasets and show that our approach outperforms state-of-the-art approaches that are pre-trained on ImageNet. Notably, we show that training our approach with multiple modalities in simulation also leads to improved object recognition on a real-world dataset (NYU-D).

II. Related Work

Our work is at the intersection between learning with privileged information for RGB-D vision and self-supervised representation learning. In the following section, we will review recent approaches from these two research fields.

A. Learning with Privileged Information

Our method follows the paradigm of learning with privileged information, which was first introduced in the seminal paper by Vapnik *et al.* [8]. They propose to use privileged (i.e., additional and valuable) information during training, by an algorithmic extension to Support-vector machines. Later, several methods for deep learning took up the idea in the context of RGB-D vision.

A recurring idea is to hallucinate features of the missing modality (typically the depth data is dropped at test time) and to combine them with the RGB features. This improves the performance of RGB object recognition and detection. Hoffman *et al.* [10] learn a hallucination network using an euclidean loss between the depth features and the hallucinated features. At test-time, the learned stream hallucinates the missing depth modality from RGB input. Garcia *et al.* [11] extend the hallucination approach with a multi-step teacher-student training. In the first stage, two individual networks (RGB and depth) are trained for the classification task. In the second stage, the weights of the depth network are frozen and the hallucination network is trained to mimic the outputs of the depth stream. In follow-up work, Garcia *et al.* [12] extend their prior method with adversarial modality distillation. All above-mentioned approaches use models pre-trained on ImageNet [13] to initialize their networks. In contrast to these works, we do not require a pre-trained ImageNet network and no additional hallucination network (which doubles the amount of parameters).

Other methods for learning with privileged information use the idea of training with the available modalities and dropping a modality during training, which forces the network to be robust against a missing modality at test-time. Neverova *et al.* [14] learn a late-fusion scheme where individual network streams are combined. The main idea is to use dropout in the fusion layer (i.e. a modality dropout). Similarly, de Blois *et al.* [15] propose input dropout. Compared to Neverova *et al.* they fuse the modalities at the input level. Specifically, the modalities are concatenated into one input tensor, where the channels of a modality are randomly dropped out (by setting the pixel values to zero) during training. Lambert *et al.* [16] propose to use privileged information to control the variance of dropout in a deep neural network and report improved performance for object classification on ImageNet.

The mentioned methods do not explicitly enforce a consistency between different modalities on a feature-level. For that reason, our method is more related to the hallucination-based approaches. For an overview on learning with privileged information and its connection to multimodal or multi-task learning we refer the reader to Jonschkowski *et al.* [17].

B. Self-supervised Representation Learning

Our method builds upon recent methods for self-supervised representation learning and multi-view learning. Specifically, we leverage recent developments in contrastive learning. Here, representations are learned by contrasting between representations of a shared context. For example, two views of the same object are used as positive and two views showing different objects are used as negative learning signal for training in a self-supervised manner. Two popular methods are Contrastive Predictive Coding (CPC) [18] and Deep InfoMax [19]. Deep InfoMax leverages the local structure in an image to learn representations. The learning signal behind the method is to classify whether a pair of global features and local features stem from the same image or not. CPC extracts spatially aligned subcrops from an image and learns an embedding for each of these crops using autoregressive models [18]. Tian *et al.* [20] extend CPC and Deep InfoMax to multiple views (e.g., depth, luminance and chrominance images) and learn a representation that is invariant to the various views (which they denote as contrastive multiview coding or CMC). Similarly, Bachman *et al.* [9] extend Deep InfoMax to the multi-view case. Compared to prior work they use strong data augmentation techniques to create new views of the same image that the representation should be invariant to. In this work, we build upon their work in order to learn with multiple modalities. We compute a set of different views from multiple sensor modalities (more specifically RGB and depth). We use the multimodal contrastive learning objective as privileged information in a multi-task learning framework. Recently, several extensions to CPC and Deep InfoMax haven been proposed that study the influence of data augmentation, simplify the contrastive loss objective and increase the model capacity [21], [22]. These methods are among the first to match the recognition performance of a supervised ResNet-50 on ImageNet.

III. Method

Our goal is to learn multimodal representations that improve the recognition performance of the model when only one modality is available at test-time. The underlying idea is that different modalities can provide complementary information during training. We assume that we have access to multimodal data for training $\mathcal{D} = \{\mathbf{x}_k^{RGB}, \mathbf{x}_k^{D}, y_k\}_{k=1}^{N}$ but during testing we only have samples from one modality (e.g., only RGB images). We use the additional modality

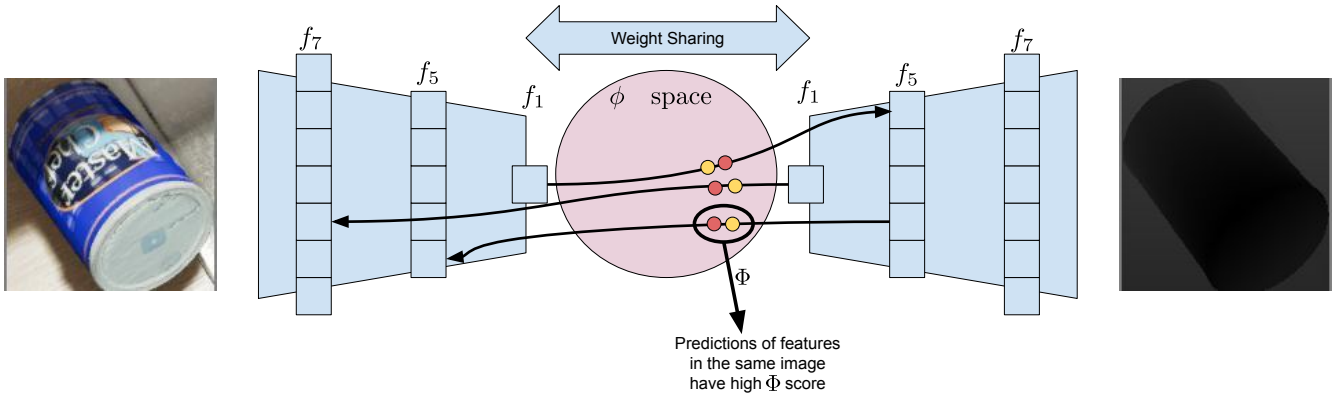

Fig. 2. The multimodal contrastive task: We learn representations that are invariant across modalities. Arrows represent contrastive objectives between features in a weight-sharing encoder. The self-supervised contrastive task is to classify whether pairs of features extracted from different modalities (x^{RGB}, x^D) depict the same concept (i.e., same object) or not.

that is available during training as privileged information. We propose to do so by optimizing a multi-task objective. The objective consists of a self-supervised contrastive learning loss and a cross-entropy classification loss. The contrastive loss captures the information between multiple sensor modalities (without using supervision) and enables to learn a representation in which similar concepts in different modalities point close to each other in the feature space. At test-time the model processes only the available modality.

A. Contrastive Loss

Contrastive objectives are used in recent self-supervised representation learning methods. The representations are learned by contrasting positive and negative examples in a self-supervised manner. Specifically, contrastive methods learn to distinguish whether a pair of global and local features stem from the same image or not. The intuition behind it is to learn which are the important signals in images and which ones are noise (e.g., background, viewpoint variations). We begin with explaining AMDIM, the self-supervised contrastive learning method we build upon, proposed by Bachman *et al.* [9].

AMDIM contrasts global features $f_1(\mathbf{x})$, extracted from the last layer, with local features $f_7(\mathbf{x})_{ij}$ extracted from a mid-level layer of a deep convolutional encoder network. The subscript $d \in \{1, 7\}$ denotes the spatial dimension $d \times d$ of the feature map (the output of a layer) and (i, j) are the indices that reference to one feature vector inside the feature map. Intuitively, AMDIM enables the global feature vector to capture only the relevant information from all the different local features.

The features are computed using sampling. First, we sample an input image $\mathbf{x} \sim \mathcal{D}$. Second, we sample spatial indices (i, j) uniformly. Third, we compute the features $f_1(\mathbf{x}), f_7(\mathbf{x})_{ij}$ for one image. A positive sample pair $(f_1(\mathbf{x}), f_7(\mathbf{x})_{ij})$ consists of global and local features from the same image, i.e., it is drawn from the joint distribution $p(f_1(\mathbf{x}), f_7(\mathbf{x})_{ij})$. A negative sample pair consists of the global feature $f_1(\mathbf{x})$ and a local feature from a different image (drawn from the marginal distribution $p(f_7(\mathbf{x})_{ij})$). In each training batch we sample a set of n_7 negative samples, N_7. The contrastive learning task is to pick the true positive pair out of a large set of negative "distractor" pairs, which results in a multi-class $(n_7 + 1)$ softmax classification loss:

$$l_\Phi(f_1, f_7, N_7) = -\log \frac{\exp(\Phi(f_1, f_7))}{\sum_{\tilde{f}_7 \in N_7 \cup \{f_7\}} \exp(\Phi(f_1, \tilde{f}_7))}, \quad (1)$$

where we omit the indices (i, j) and the dependence between f and $\mathbf{x}$ to shorten the notation. The denominator term consists of one positive sample pair and n_7 negative sample pairs. The function $\Phi(f_1, f_7)$ is a scoring function that maps feature pairs to a scalar value (high for positive samples and vice versa). Specifically, the function that we approximate takes as input two features $(f_1(\mathbf{x}), f_7(\mathbf{x}))$ and computes a scalar matching score Φ. Specifically, we compute the dot product between the output of two single-layer networks ϕ_1 and ϕ_7 as follows:

$$\Phi(f_1(\mathbf{x}), f_7(\mathbf{x})_{ij}) \triangleq \phi_1(f_1(\mathbf{x}))^\top \phi_7(f_7(\mathbf{x})_{ij}). \quad (2)$$

B. Multimodal Contrastive Loss

Our method extends the AMDIM approach to multiple modalities. In addition we incorporate data augmentation that generalizes to both modalities RGB and depth. Specifically, we sample the multimodal features as follows: First, we sample an image pair from our multimodal dataset $(\mathbf{x}^{RGB}, \mathbf{x}^D) \sim \mathcal{D}$. Next, we augment each image from a modality $M \in \{RGB, D\}$ with a modality-dependent augmentation function

$$\tilde{\mathbf{x}}^M \sim \mathcal{A}^M(\mathbf{x}^M). \quad (3)$$

We use RandAugment [23] to construct modality-specific augmentation functions, which apply a set of randomly sampled augmentations to an image. Then, we sample spatial indices (i, j) and compute the features for all augmented modalities $f_1(\tilde{\mathbf{x}}^{RGB})$, $f_1(\tilde{\mathbf{x}}^D)$, $f_7(\tilde{\mathbf{x}}^{RGB})_{i,j}$ and $f_7(\tilde{\mathbf{x}}^D)_{i,j}$. The joint distributions are $p^{\mathcal{A}}(f_1(\tilde{\mathbf{x}}^{RGB}), f_7(\tilde{\mathbf{x}}^D)_{ij})$ and $p^{\mathcal{A}}(f_1(\tilde{\mathbf{x}}^D), f_7(\tilde{\mathbf{x}}^{RGB})_{ij})$. The marginal distributions that

we use for sampling negatives are $p^{\mathcal{A}}(f_7(\tilde{\mathbf{x}}^{RGB})_{ij})$ and $p^{\mathcal{A}}(f_7(\tilde{\mathbf{x}}^{D})_{ij})$. The two resulting multimodal contrastive objectives look as follows:

$$\mathcal{L}_{\Phi}^{RGB\to D} = \mathop{\mathbb{E}}_{(f_1(\tilde{\mathbf{x}}^{RGB}),f_7(\tilde{\mathbf{x}}^{D})_{ij})} \left[\mathop{\mathbb{E}}_{N_7} \left[l_{\Phi}(f_1, f_7, N_7) \right] \right], \quad (4)$$

$$\mathcal{L}_{\Phi}^{D\to RGB} = \mathop{\mathbb{E}}_{(f_1(\tilde{\mathbf{x}}^{D}),f_7(\tilde{\mathbf{x}}^{RGB})_{ij})} \left[\mathop{\mathbb{E}}_{N_7} \left[l_{\Phi}(f_1, f_7, N_7) \right] \right]. \quad (5)$$

The expectations are approximated using the samples in each training batch. We also adopt the multiscale formulation of the cost proposed in AMDIM but omit the formula to ease reading. Specifically, in our mathematical formulation we only use two features (f_1 and f_7) but in practice we use three feature levels (f_1, f_5, f_7) to compute the contrastive loss from f_1-to-f_5, f_1-to-f_7, and f_5-to-f_5. In Figure 2 we show the multiscale connections between the various feature levels.

C. Multitask learning loss

To optimize our final objective we combine the described contrastive loss (CL) with a cross-entropy loss $\mathcal{L}_{class}^{M}$. The cross-entropy loss (CE) in our case is unimodal (i.e., we forward the feature activations from one modality to the classifier). The classifier is a fully-connected neural network that we add to the last layer of the encoder. The gradient of the classification layer is back-propagated through the encoder and therefore our method is a single-step end-to-end approach. The final objective looks as follows:

$$\mathcal{L}_{total}^{M} = \mathcal{L}_{\Phi}^{RGB\to D} + \mathcal{L}_{\Phi}^{D\to RGB} + \mathcal{L}_{class}^{M}. \quad (6)$$

D. Hyperparameters

In order to train the encoder network on one NVIDIA Titan-X GPU we modify the original convolutional encoder proposed by Bachman *et al.*. We set the encoder feature dimension of the last layer to $ndf = 128$. We also reduce the output dimension of the embedding function $\phi_d(f_d)$ to $nrkhs = 1024$. In addition, we set the depth of the residual blocks to $ndepth = 8$. Overall, this reduces the number of learnable parameters from 696 million to 93.7 million parameters. Therefore our model is comparable to a ResNet 50 (2x) as used in Chen *et al.* [22] that has 94 million parameters. In all experiments we use 128×128 as size for the input images.

Algorithm 1 Encoder architecture adapted from Bachman *et al.* [9]

ReLU(Conv2d(3, ndf, 5, 2, 2))
ReLU(Conv2d(ndf, ndf, 5, 2, 2))
ResBlock(1*ndf, 2*ndf, 4, 2, ndepth)
ResBlock(2*ndf, 4*ndf, 4, 2, ndepth)
ResBlock(4*ndf, 8*ndf, 2, 2, ndepth) {extract f_7}
ResBlock(8*ndf, 8*ndf, 3, 1, ndepth) {extract f_5}
ResBlock(8*ndf, 8*ndf, 3, 1, ndepth)
ResBlock(8*ndf, nrkhs, 3, 1, 1) {extract f_1}

IV. Experiments

In the following section, we will test our method through extensive experiments. Our aim is to show that we can improve the performance in the unimodal case if we train with multimodal data. First, we test our method on the Falling Things dataset and compare against CMC [20], a very recent contrastive learning method. Then we evaluate our method on the NYU RGB-D dataset and show how we can further improve the results by transfer learning from Falling Things. To conclude our evaluation, we investigate the performance of our method in a semi-supervised setting.

A. Falling Things RGB-D

The Falling Things (FAT) dataset from NVIDIA [24] provides rendered images with artificially placed 3D household object models in virtual environments. For these scenes 3D poses, per-pixel class segmentations, and 2D/3D bounding-box coordinates are provided for each placed object. The dataset contains three modalities, mono RGB, stereo RGB and depth images. We use the provided bounding box information to crop the individual objects from the scene for constructing a classification dataset. This results in $18,640$ training images and $6,267$ validation images. Each image contains an object out of 21 household objects from the YCB dataset [25]. In order to process the corresponding inputs from different modalities with a single encoder, the input dimensions have to match. To match RGB and depth dimensions, we stack the one-dimensional depth image three times. We train all models from scratch and do not use any pre-trained models on this dataset.

We first analyze the results when training with RGB-D compared to training with a single modality. All results can be found in Table I. Choosing the RGB domain as target (=test) modality, we first train the custom ResNet (ndf=128, nrkhs=1024, ndepth=8) with the cross-entropy classification loss in a fully-supervised manner. The fully supervised network achieves 92.5% accuracy. Next, we evaluate the performance for combining the cross-entropy loss with the contrastive loss, again in the unimodal case. Here, we achieve a similar performance of 92.3% after 130 epochs of training. Next, we add the second modality during training. First, we evaluate using the original two-step training procedure of AMDIM; self-supervised pre-training of the encoder followed by a supervised training of a classifier on top of the fixed encoder. This yields a classification accuracy of 91.1%. Finally, we train our method that uses multiple modalities and the multi-task loss. We achieve a classification accuracy of 94.4%. This is an improvement of 2.1% in comparison to the best unimodal model.

Next, we study the effect of stronger data augmentation, specifically when using RandAugment (RA) [23]. The classification accuracy in the unimodal setting (RGB) slightly improves to 95.1%. In the multimodal setting we see an improvement of 0.7% to 95.1% accuracy.

We perform the same set of experiments but now we switch the test modality to depth. The fully supervised baseline achieves a classification accuracy of 79.2%. Adding

TABLE I

CLASSIFICATION ACCURACY ON THE FALLING-THINGS DATASET. WE REPORT THE TRAINING AND TESTING MODALITY, THE ACTIVATION OF THE LOSS TERMS, THE NUMBER OF EPOCHS, AND THE TOP-1 CLASSIFICATION ACCURACY FOR ONE TRAINING RUN.

Ours						
Train modality		Test modality	CL	CE	Epochs	Top-1 Acc.
RGB		RGB	-	✓	100	92.5%
RGB		RGB	✓	✓	130	92.3%
RGB	Depth	RGB	$\checkmark^{1st}$	$\checkmark^{2nd}$	100+30	91.1%
RGB	Depth	RGB	✓	✓	130	94.4%
RGB+RA		RGB	✓	✓	130	92.6%
RGB+RA	Depth+RA	RGB	✓	✓	130	95.1%
Depth		Depth	-	✓	100	79.2%
Depth		Depth	✓	✓	130	85.1%
RGB	Depth	Depth	$\checkmark^{1st}$	$\checkmark^{2nd}$	100+30	84.1%
RGB	Depth	Depth	✓	✓	130	86.1%
Depth+RA		Depth	✓	✓	130	85.4%
RGB+RA	Depth+RA	Depth	✓	✓	130	91.6%
CMC						
Unsupervised RGB-D, Supervised RGB					100+30	81.2%
Unsupervised RGB-D, Supervised D					100+30	56.8%
Unsupervised RGB-D, Supervised RGB					400+100	86.28%
Unsupervised RGB-D, Supervised D					400+100	73.82%

the contrastive loss term improves the performance to 85.1%. Adding the second modality and training using the original two-step procedure does not improve the results (84.1% classification accuracy). Training with our multi-task loss in an end-to-end manner improves the accuracy to 86.1%. We achieve the best results when adding RA to our approach. The classification accuracy reaches 91.6%, which is an improvement of 6.5% with respect to the best unimodal baseline (trained with RandAugment as well).

Finally, we compare our method against contrastive multiview coding (CMC). We use the publicly-available code and do not make changes to CMC. We train the encoder for 100 epochs in an unsupervised fashion and then the classifier using the learned representation for another 30 epochs. We keep the hyperparameters in this setting fixed and use the suggested parameters of CMC for training on ImageNet with the ResNet101 as encoder. To compare CMC to our method we train it with two modalities (RGB and depth). When testing on RGB images CMC achieves a performance of 81.2%. For the depth modality the performance is 56.8%. We can see that our approach outperforms CMC in all cases. When training according to the full schedule of 400 unsupervied epochs and 100 supervised epochs, we achieve an depth accuracy of 86.28% and a

Discussion: We show that our multimodal multi-task loss, optimized in a single stage, improves the performance (when testing with RGB) whereas the original two-step training approach of AMDIM does not. The data augmentation only slightly improves the results, likely because on the RGB modality the performance on this instance recognition task is already high. Contrary, when training with RGB-D and testing on depth, adding data augmentation improves the results by 5.5%.

B. NYU-D

As second dataset we use NYU-D which was originally proposed by Silberman *et al.* [26]. We use a variant of this RGB-D dataset, which was proposed in [27] for object classification. This variant crops tight bounding boxes around instances of 19 object classes that are present in the dataset. The resulting dataset consists of 2,186 paired and labeled training images and 2,401 test images. In this dataset the depth images are color-encoded using the HHA encoding [28].

TABLE II

CLASSIFICATION ACCURACY ON THE NYU RGB-D DATASET. WE REPORT THE TRAINING AND TESTING MODALITY AND THE TOP-1 CLASSIFICATION ACCURACY FOR ONE TRAINING RUN. WE TRAIN USING OUR MULTI-TASK LOSS FOR 130 EPOCHS WITH A BATCH-SIZE OF 40. ⋆ = TRAINED FOR 260 EPOCHS

Train Modality		Test Modality	Top-1 Acc.
RGB		RGB	37.8%
RGB	Depth	RGB	42.6%
RGB+RA		RGB	46.4%
RGB+RA	Depth+RA	RGB	49.4%
RGB+RA	Depth+RA	RGB	50.1%⋆
Depth		Depth	55.2%
RGB	Depth	Depth	55.1%
Depth+RA		Depth	55.5%
RGB+RA	Depth+RA	Depth	57.9%

All models are trained with the same hyper parameters as on Falling Things. However, the batch-size is reduced to 40 in order to have a more comparable number of gradient steps given that the dataset contains less images. We report the results after 130 epochs of training. We train all models from scratch.

We report results for the same two scenarios; testing on depth and testing on RGB, see Table II. Unimodal training yields a performance of 37.8% on the RGB testset. Multimodal training improves the result to 42.6%. When we add RandAugment in the unimodal case the results improve to 46.4%. Combining RGB and Depth together with RandAugment improves the result to 49.4%.

Next, we investigate testing on depth as the target modality. First, we report small gains between unimodal and multimodal training (55.1% compared to 55.2%). However, adding RandAugment to our methods benefits the classification of depth data at test-time. Notably, in this setting the results improve by 2.4% to 57.9%.

We can conclude that training with two modalities consistently improves the results in combination with data augmentation.

C. Transfer Learning: Falling Things to NYU-D

In our prior experiments, we did not perform pre-training of the encoder on ImageNet or on other datasets. In this section, we will investigate whether transfer learning our best Falling Things model improves performance on NYU-D. Our goal is to show that we can outperform methods pre-trained on ImageNet by pre-training on Falling Things, which notably is a synthetic dataset.

For all experiments, we initialize the encoder with the weights of a Falling Things model that we trained in previous experiments. Then, we fine-tune the model for another 130

epochs with a batch-size of 40. We run the fine-tuning of the final model multiple times (following the literature) to estimate the variance of our approach.

First, we take the best performing network on the Falling Things dataset on the (RGB, Depth → RGB) and the (RGB, Depth → RGB) task. In both cases, we see that pre-training on FAT helps and improves the classification performance by 2% with an average performance of 51.92% or 52.0%, see Table III. Both pre-trained feature extractors were trained on FAT with our multi-task loss, i.e., class information from FAT propagates into the encoder.

For that reason, we fine-tune a model that was pre-trained in an unsupervised manner on FAT, i.e., this model was trained only with $\mathcal{L}_\Phi$. However, for fine-tuning on NYU-D we use our multi-task loss again. Interestingly, fine-tuning from the unsupervised encoder achieves the best accuracy with 53.84%. This is an improvement of about 4% compared to using no pre-training on FAT. For this run, we report the results across 12 different runs, with two different encoders trained in the same unsupervised fashion on FAT.

We compare our best model to four state-of-the-art methods from the privileged information literature that learn with multimodal data and test on unimodal data.

ModDrop [14]: the method trains several networks for each modality. The networks are combined in a late-fusion layer, which is trained using dropout. Due to training with modality dropout, the method can also be tested with a single modality.

ADMD [12]: the method uses a hallucination network that generates features of the missing modality at test-time. The hallucination is learned using a generative adversarial network.

Input Dropout [15]: the method concatenates all modalities during training in one input tensor. During training one modality is randomly dropped out of the input tensor (pixels are set to zero). The network learns to ignore the missing modality at test time.

Pix2Pix GAN [29]: the method uses a generative adversarial network to hallucinate depth images from RGB input. In contrast to ADMD it hallucinates depth images instead of high-level depth features.

Table III depicts how our method quantitatively compares to these methods. Our method outperforms all other methods when averaging the results over five training runs. We outperform the best-performing method of Garcia *et al.* (ADMD) by 2.5%. We discuss the details on this in subsection IV-D. Finally, we emphasize that our method is the only one not pre-trained on ImageNet but only on synthetic data.

Discussion: This effect can not be attributed to the longer training, because when training 260 epochs from scratch we only can see a mild performance increase to 50.1% (see Table II). We can conclude that pre-training is beneficial for our method, even when using only synthetic data. This is especially interesting for cases in which real-world training data is scarse. We show that it is possible to pre-train a network on synthetic data in an unsupervised fashion and then use this network for fine-tuning on a smaller real-world dataset such as NYU-D.

TABLE III

TRANSFER LEARNING RESULTS: SYNTHETIC DATA (FALLING THINGS, SCENENET) TO NYU RGB-D. ALL MODELS ARE FINE-TUNED WITH A BATCH-SIZE OF 40 AND FOR 130 EPOCHS ON NYU RGB-D. N INDICATES HOW MANY INDIVIDUAL RUNS HAVE BEEN TRAINED.

Train Modality		Test Modality	Top-1	Std.	N	Source
RGB+RA	Depth+RA	RGB	51.92%	0.63%	5	Our[1]
RGB+RA	Depth+RA	RGB	52.00%	0.88%	5	Our[2]
RGB+RA	Depth+RA	RGB	53.84%	0.77%	12	Our[3]
RGB + RA	Depth + RA	RGB	59.92%	0.59%	5	Our[4]
Baselines						
ModDrop		RGB	44.3%	n/a⋆	5	[15]
Pix2Pix		RGB	48.2%	n/a⋆	5	[15]
Input Dropout		RGB	49.5%	0.80⋆	5	[15]
Input Dropout + RGB		RGB	52.7%	0.60⋆	5	[15]
Mod. distillation (ADMD)		RGB	57.4%	0.3⋆	5	code by [12]

⋆ provided by author, some models not available anymore
[1] best performing model on the RGB-Depth → RGB on Falling Things
[2] best performing model on the RGB-Depth → Depth on Falling Things
[3] feature extractor trained unsupervised on Falling Things RGB-D
[4] feature extractor trained unsupervised on SceneNet RGB-D

D. Transfer Learning: SceneNet to NYU-D

We perform an additional transfer learning experiment using SceneNet RGB-D, a photorealistic dataset of 5M synthetic indoor images, for pre-training. We construct our classification subset for training by cropping objects from the synthetic scenes using the provided bounding-box annotations. We use a smaller subset of the original data and build a dataset that contains 202,747 training and 33,574 validation images containing 16 classes. Note that the classes partly overlap with the NYU-D dataset but as before we do not use the class information in our unsupervised pre-training. Results in Table III show that pre-training on SceneNet leads to an improvement in classification performance of 6% compared to pre-training on the Falling Things dataset. We attribute this improvement to the larger training set size and the smaller domain mismatch.

E. Ablation: Augmentation for Depth Modality

The introduction of RandAugment [23] as primary source for data augmentation has yield great benefits on both datasets. Since RandAugment was initially designed for RGB images, it is was unclear how to augment the depth modality. Therefore, we examined which set of augmentations of RandAugment can be re-used for depth data.

As done in [23], we evaluate each of the individual augmentations proposed in RandAugment and how they affect the classification accuracy when training and testing on the depth modality. The results of this experiments can be found in Table IV. All experiments have been conducted on the NYU RGB-D dataset with 100 epochs of training and a batch-size of 140.

From this experiment we can see that all the augmentations which directly change values of pixels within the depth map lead to a vast reduction in classification accuracy. For that reason, we remove all these augmentations from the set of possible augmentations. We then evaluate two sets of augmentations. The first contains all augmentations which do not harm the classification accuracy by more than one percentage point. The second one contains the two strongest

TABLE IV

COMPARISON OF RESULTS FOR THE NYU RGB-D DATASET. EACH MODEL IS TRAINED FOR 100 EPOCHS WITH BATCH-SIZE 140.

Modality	Augmentation	Test modality	Acc. (TOP-1)	Change
Depth	-	Depth	55.3%	-
Depth	Cutout	Depth	55.8%	+0.5%
Depth	Rotate	Depth	56.2%	+0.9%
Depth	Translate-X	Depth	55.8%	+0.5%
Depth	Translate-Y	Depth	54.1%	−1.2%
Depth	Shear-X	Depth	55.2%	−0.1%
Depth	Shear-Y	Depth	55.7%	+0.4%
Depth	Sharpness	Depth	55.2%	−0.1%
Depth	Brightness	Depth	16.3%	−39.0%
Depth	Contrast	Depth	18.4%	−36.9%
Depth	Invert	Depth	14.0%	−41.3%
Depth	Equalize	Depth	29.8%	−25.5%
Depth	Auto-Contrast	Depth	20.9%	−34.4%
Depth	Speckle	Depth	11.1%	−44.2%
Depth	RandAugment[1]	Depth	55.5%	+0.2%
Depth	RandAugment[2]	Depth	55.7%	+0.4%
1	this set of augmentations contains: Cutout, Rotate, Translate-X, Shear-X, Shear-Y, and Sharpness			
2	this set of augmentations contains: Cutout, Rotate			

augmentations on the depth modality. Both set of augmentations perform somewhat similar with 55.5% and 55.7% classification accuracy. Nevertheless, we used the latter set of augmentations in our previously reported main experiments.

Discussion: Data augmentation for depth modality remains an open question. In this experiment, we can see that all augmentations that directly change depth values do not perform well. The reason for that is that they change the perceived objects quite drastically. More elaborate augmentations which change the viewpoint of the camera, keep the scene geometry and create a new depth image from different viewpoints might improve results.

F. Ablation: Semi-Supervised

In the following, we evaluate our method in a semi-supervised setting on Falling Things. We manually vary the amount of labeled data and use the remaining data for the unsupervised contrastive loss. We train our method on RGB-D and test on depth. We compare against unimodal training and testing. In this experiment we see that training with multiple modalities is beneficial, especially in the part of the curve where less labels are provided, see Figure 3. Our multimodal method achieves with 10% labeled data a classification accuracy of 73.7%, compared to the 55.5% of the unimodal baseline. With more labeled data we see the advantage of multimodal training reducing.

Discussion: This experiment shows the advantage of using the second modality when small amounts of labeled data is available. This advantage is amplified by the fact that we did not use data augmentation to further regularize the learning of the encoder. These results also fall in line with research by Henaff *et al.* [21] who also find that a contrastive loss function performs well in the semi-supervised setting.

V. CONCLUSION

We introduced a novel approach for learning with privileged information. Following this approach, we developed a novel method that leverages multimodal contrastive learning

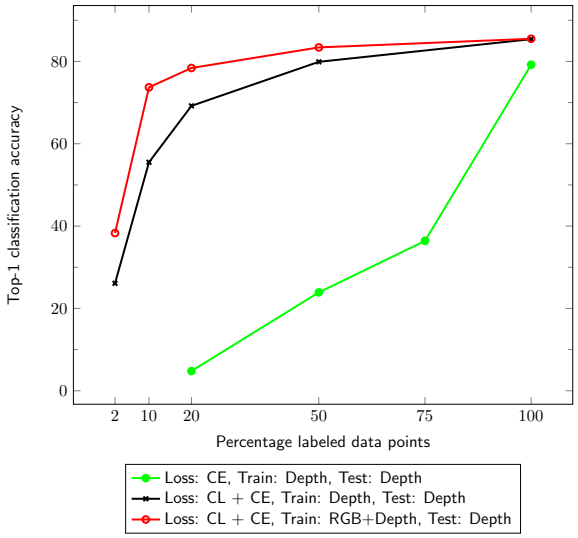

Fig. 3. Top-1 classification accuracy with different percentages of labeled data provided. The target modality in this setting is depth. The red line shows our method trained with RGB-D and the multi-task loss. The black curve shows our method in the unimodal case (trained only on depth) and the green curve shows the same encoder trained only with a cross-entropy loss.

to improve unimodal object recognition. We showed in extensive experiments that we can learn rich representations from multiple modalities that improve the discriminative performance in scenarios where only one modality is available. Future work could extend our method to improve unimodal object detectors or semantic segmentation networks with multimodal training. In addition, training our method with a large amount of sensor modalities in simulation seems a promising research direction for Sim2Real tasks. We believe that learning with privileged information in multimodal settings is an important direction towards improving robot vision.

REFERENCES

[1] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Unsupervised Feature Learning for RGB-D Based Object Recognition. In Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar, editors, *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, Springer Tracts in Advanced Robotics, pages 387–402. Springer International Publishing, Heidelberg, 2013.

[2] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.

[3] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018.

[4] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-Supervised Model Adaptation for Multimodal Semantic Segmentation. *International Journal of Computer Vision*, July 2019.

[5] O. Mees, A. Eitel, and W. Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156, Oct 2016.

[6] Inkyu Sa, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. Deepfruits: A fruit detection system using deep neural networks. *Sensors*, 16(8):1222, 2016.

[7] Guan-Horng Liu, Avinash Siravuru, Sai Prabhakar, Manuela Veloso, and George Kantor. Learning end-to-end multimodal sensor policies for autonomous navigation. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 249–261. PMLR, 13–15 Nov 2017.

[8] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. In *2009 International Joint Conference on Neural Networks (IJCNN2009)*, volume 22, pages 544–557, 2009.

[9] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 15509–15519, 2019.

[10] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, June 2016.

[11] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–121, 2018.

[12] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 60(6):84–90, 2017.

[14] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[15] Sébastien de Blois, Mathieu Garon, Christian Gagné, and Jean-François Lalonde. Input dropout for spatially aligned modalities. *arXiv preprint arXiv:2002.02852*, 2020.

[16] J. Lambert, O. Sener, and S. Savarese. Deep learning under privileged information using heteroscedastic dropout. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, June 2018.

[17] Rico Jonschkowski, Sebastian Höfer, and Oliver Brock. Patterns for learning with side information. *arXiv preprint arXiv:1511.06429*, 2015.

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

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[20] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[21] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[23] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical data augmentation with no separate search. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[24] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2038–2041, 2018.

[25] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, Sep. 2015.

[26] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[27] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[28] Saurabh Gupta, Ross Girshick, Pablo Arbelaez, and Jitendra Malik. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014*, Lecture Notes in Computer Science, pages 345–360, Cham, 2014. Springer International Publishing.

[29] B. Bischke, P. Helber, F. Koenig, D. Borth, and A. Dengel. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, Sep. 2018.