Cross-Modal Adaptation for RGB-D Detection

Judy Hoffman[1] and Saurabh Gupta[1] and Jian Leong[1] and Sergio Guadarrama[2] and Trevor Darrell[1]

***Abstract*— In this paper we propose a technique to adapt convolutional neural network (CNN) based object detectors trained on RGB images to effectively leverage depth images at test time to boost detection performance. Given labeled depth images for a handful of categories we adapt an RGB object detector for a new category such that it can now use depth images in addition to RGB images at test time to produce more accurate detections. Our approach is built upon the observation that lower layers of a CNN are largely task and category agnostic and domain specific while higher layers are largely task and category specific while being domain agnostic. We operationalize this observation by proposing a mid-level fusion of RGB and depth CNNs. Experimental evaluation on the challenging NYUD2 dataset shows that our proposed adaptation technique results in an average 21% relative improvement in detection performance over an RGB-only baseline even when no depth training data is available for the particular category evaluated. We believe our proposed technique will extend advances made in computer vision to RGB-D data leading to improvements in performance at little additional annotation effort.**

I. Introduction

Accurate object detection is an essential component for many robotic tasks like mapping, motion planning, grasping and object manipulation. This has motivated the use of depth information from commodity RGB-D sensors to improve object recognition performance [20], [19], [32], [31], [47].

However, most well performing methods rely on Convolutional Neural Networks (CNNs) to learn features for depth images and require a large amount of annotated examples to be effective. Numerous efforts in the vision community over the last 15 years have led to the development of large scale RGB datasets [9], [12], [35], which have enabled huge progress on a variety of problems. However, while labeled RGB data is currently available for hundreds of categories with strong annotations and for thousands with weak annotations, the available labeled depth data is currently limited to tens of categories.

At the same time, the introduction of low cost and easy to use RGB-D image capturing systems has enabled many robotic setups to have access to both RGB and depth information during operation. Current techniques require bounding box annotations to train object detectors and limit use of depth images to categories for which such annotations exist. Thus, even though a depth sensor is available at test time, researchers are forced to use RGB-only detectors for most object categories they may want to study. This

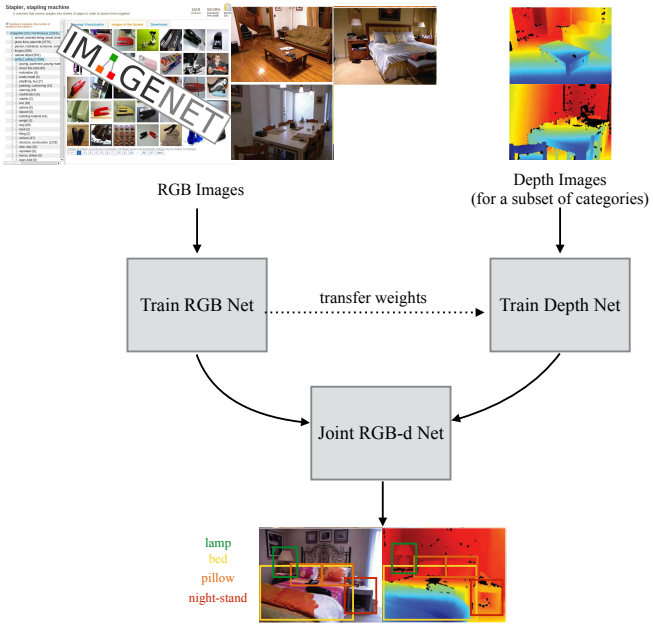

Fig. 1. Given labeled depth images for a handful of categories we adapt an RGB object detector for a new category such that it can now use depth images in addition to RGB images at test time to produce more accurate detections. We do this by fusing information across modalities and use the available labeled depth data to extract mid-level depth representations which can be processed into semantic class labels for improved test time recognition performance on all categories of interest.

situation presents us with an interesting question: are detailed bounding box annotations for all object categories necessary to enable improved test time recognition using additional modalities. Or is there a way to utilize the vast amounts of labeled RGB data already available, along with limited labeled depth data, to train object detectors which can use RGB-D images at test time to boost performance over an RGB detector, even for objects with no labeled depth examples?

In this work, we address this question and propose a transfer approach which leverages labeled RGB-D data for some categories (denoted as auxiliary categories) to build RGB-D object detectors for additional categories for which we only have RGB training data (see Figure 1). We do this by fusing mid-level representations from depth and RGB images. This fused mid-level representation can be used with RGB-only classifiers to improve the quality of the RGB detector.

We evaluate our technique on the challenging NYUD2 dataset and our experiments show that we are able to

[1]Department of Electrical Engineering and Computer Science at University of California, Berkeley. {jhoffman, sgupta, trevor}@eecs.berkeley.edu, jianwei.leong@berkeley.edu

[2]Google Research, Mountain View. sguada@gmail.com

effectively adapt RGB detectors into RGB-D detectors. These RGB-D detectors can effectively leverage depth data at test time and we observe a 21% relative improvement over an RGB-only detector. Note that this was done without using any depth training data for the evaluated categories. We believe that our technique will facilitate the transfer of progress made in computer vision to fields like robotics.

II. Related Work

We review three major bodies of research relevant to our work here, multi-modal and multi-domain adaptation techniques, techniques for generating region proposals and object detection with RGB-D images.

a) Transferring Information Across Tasks: Many methods have been proposed to transfer general information between different data sources for related tasks [39], [30], [17], [23], [14], [6]. Multi-modal deep learning architectures have been explored previously in a generative context [36], [44], and parallel convnet architectures have been previously explored in the context of Siamese network learning [5], [7]. Given the ease of collecting annotations for an image classification task, as opposed to an object detection task, there have been many techniques proposed to train detectors from weak labels [41], [2], [1], [50]. These method are notoriously hard to optimize and must be trained independently for each detection category. A recent method was proposed to transfer generic information from CNN based detectors to transform CNN classifiers into object detectors [24]. Although effective, it was limited to transferring information between RGB models. Other approaches have been proposed to transfer generic information across modalities [8], but have only been shown with weak detection models.

b) Region Proposals: We note that many top-performing supervised object detection methods [15] and weakly supervised methods [41], [24] rely on a good set of bottom-up bounding box object candidates. Object proposal generation has been an active area of research in computer vision in recent years [3], [28], [52], [49]. Given the importance of good region proposals [25], naturally people have studied the problem of using depth images to improve the quality of object proposals [34], [20]. Gupta *et. al.*[20] use depth information to obtain improved contours from RGB-D images, and use this in a multi-scale combinatorial grouping framework [3] to report great improvements over RGB only methods, obtaining the same recall with an order of magnitude fewer regions as compared to RGB only methods.

c) Object Detection: Lastly, there has been considerable work on the problem of object detection for RGB-D images [26], [43], [32], [48], [51], [20], [42], [31]. [26], [43], [48], [51] propose extensions to deformable part based models [13] to compute additional features from the depth image, and report performance improvements over just using the RGB image. Song and Xiao [42] design rich features on the depth images while Gupta *et. al.*[20] proposed a novel geocentric embedding for learning features from depth images, and both these methods report great improvements over previous works. While all of these methods report significant improvements over RGB-only methods, they all require bounding box annotations to train their models. In our work, we build off the ideas from LSDA [24] to allow us to adapt a CNN model trained for one task, which has plentiful training data, to perform a different test time task which has limited training data.

III. Method

In this section, we describe our method for learning object detection models that use depth information from auxiliary categories to improve test-time performance for a new category.

We use $\mathcal{L}$ to denote the set of auxiliary categories for which we have annotated RGB-D data (bounding boxes around instances of the object in RGB-D scenes). We use $\mathcal{U}$ to denote the set of categories for which we only have labeled RGB data (again bounding boxes around instances of the object in RGB scenes). Our goal is to leverage depth representations learnt by training RGB-D detectors for auxiliary categories $\mathcal{L}$ to adapt RGB object detectors for categories $\mathcal{U}$ to RGB-D input, that is they can now start using RGB-D input and potentially generate better output.

Intuitively, our method uses labeled depth training data for auxiliary categories $\mathcal{L}$ to learn a mid-level representation for depth images, which can be combined with mid-level representation from RGB images at test time. This mid-level fusion of representations can be used to adapt and improve a RGB object detector for the set of categories $\mathcal{U}$. The resulting RGB-D detector is able to utilize the depth data provided at test time to improve detection, without ever being trained on any depth data for categories $\mathcal{U}$.

Most state-of-the-art object detection models follow a two stage approach:

1) Computing region proposals: These are bounding boxes on the image which have high overlap with objects in the image.
2) Scoring region proposals: This is typically done by using CNNs [15], [22], [37], [46]. CNNs learn hierarchical feature representations in an end-to-end manner.

Our proposed technique incorporates depth information into both stages of this pipeline. For region proposals, we experimented with an adaptation of Edge Boxes [52] to depth images and RGB-D MCG [20]. We found RGB-D MCG to perform better and hence use these.

Next, we describe our technique for training multi-modal CNN based architectures with incomplete training data from one modality. In our case, we have complete RGB training data and limited depth training data.

A. Incorporating Depth into the CNN Representation

Our key insight is to fuse representations from RGB and depth images at an appropriate mid-level. Given a pair of RGB and depth images of a scene, the visual concepts depicted in both images are the same, though the pixel values may differ significantly. This motivates a processing pipeline which allows independent domain specific processing to arrive at a common mid-level representation, which can then be

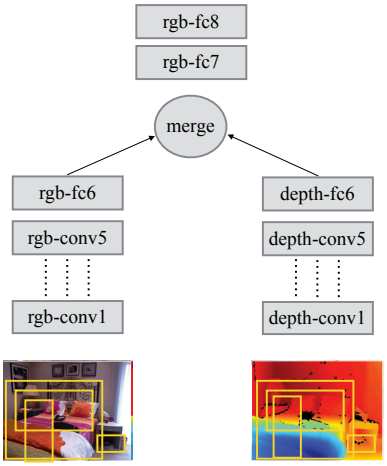

Fig. 2. Our CNN architecture. We have parallel modality-specific lower layers and merge the two branches at a semantically meaningful higher layer.

processed domain agnostically to obtain the desired semantic output. Thus, the domain specific learning can happen in the lower layers. These lower layers are often category agnostic (but domain specific) and can be trained effectively using data from a small set of categories, and can then be used with category specific but domain agnostic higher layers trained in a different domain or modality. Recent work on analyzing CNN architectures [33] in-fact shows quantitative evidence towards domain specific lower layers and task or category specific higher layers. To operationalize these findings, we use labeled RGB-D data from categories $\mathcal{L}$ to learn the domain specific but category independent lower layers and we use category specific but domain agnostic higher layers to obtain detectors for categories which lack labeled data in one of the modalities ($\mathcal{U}$).

Our proposed multi-modal architecture is depicted in Figure 2. We work with the popular AlexNet architecture [29]. AlexNet has five convolutional layers, three max pooling layers, and three fully connected layers. We use this architecture as a starting point for both the RGB and depth branches. Our insights about mid-level fusion and our training procedure are independent of the base CNN and should naturally extend to other CNN architectures.

It has been shown that the activations from layers fc6 and fc7 (the fully connected layers) produce semantically meaningful embeddings [11], [4]. We thus experimented with various fuse points in the fully connected layers, and found that fusing at fc6 worked better than both spatial fusion at pool5 and late fusion after fc7 (Section IV-B). For fusion we average the fc6 activations, after relu, of both branches and connect them with the 4096-dimensional fc7 layer, which is in turn connected to our final fc8 classifiers. We experimented with both average and concatenation as fusion techniques and found average to be slightly more robust.

B. Sequential Fine-Tuning

With the network structure determined, we now describe our method for training the network parameters. Since we lack depth training data for all categories in $\mathcal{U}$, we cannot naïvely fine-tune the full network. Instead, we propose a sequential fine-tuning procedure whereby the parameters of the RGB and depth networks are learned independently using all available labeled data from each modality.

Our training procedure is illustrated in Figure 1. We begin by training an RGB network (with AlexNet architecture), using labeled RGB data from all categories ($\mathcal{U} \cup \mathcal{L}$). We follow the standard practice of initializing this network from one that was pre-trained on the ImageNet dataset [9] for the task of image classification [11].

Next, we would like to produce an identical architecture that uses depth input in the form of an HHA encoding [20] (which encodes a depth image geocentrically using three channels: horizontal disparity, height above ground, and angle between the pixel's local surface normal and the inferred gravity direction). However, since we only have depth training data for categories in $\mathcal{L}$, we can not fine-tune the network from scratch.

Instead, we begin by populating all the weights of our depth network using the fully trained weights of our RGB network. By doing so, we initialize our depth network with parameters which have been tuned to perform well on all categories of interest, and in particular categories for which there is no depth training data. Additionally, initializing the depth network with RGB weights enables a favorable alignment between the two networks so they may be effectively combined later.

We next fine-tune the depth network on all available depth training data, allowing it to adapt to the new depth modality. Fine-tuning from RGB to depth HHA images is possible because the two modalities have similar structures [20] and higher level semantic information (e.g. object boundary information) is present in both.

Finally, after both the RGB and depth networks have been fine-tuned, we produce the final multi-modal network parameter values. For layers before the merge point, we transfer the weights from the RGB and depth networks directly to the corresponding weights of our architecture. For all layer weights above the merge point, we use the RGB model weights. This corresponds to reversing the upper depth weights back to their initialization point. We do this since the RGB parameters were learned using all labels for the portion of the model which processes mid-level representations into the final semantic outputs as opposed to the trained depth layers which have no recognition of the held out categories in $\mathcal{U}$.

IV. Experiments

A. Dataset and Setup

We evaluate our algorithm with the NYUD2 dataset [40], using the standard split of 795 training images and 654 testing images. The split is selected such that images from

the same scene do not co-occur in both sets. For all our experiments, we use annotations of the 19 major furniture categories: bathtub, bed, bookshelf, box, chair, counter, desk, door, dresser, garbage-bin, lamp, monitor, night-stand, pillow, sink, sofa, table, television, and toilet.

For all algorithms we use RGB-D MCG proposals [20]. MCG [3] generates a multi-scale hierarchical segmentation which is then used to generate region proposals. The proposals are then ranked by random forest regressors trained on features computed from the image and the region shape. Gupta *et. al.*[20] generalized this to RGB-D images by using improved edge maps [20], [10], [18] and using features from the depth image in addition to features from the RGB image and the region shape for re-ranking the proposals. RGB-D MCG produces state-of-the-art region proposals for RGB-D images and we use these for our experiments.

In addition, all variants of our algorithm as well as all baseline and state-of-the-art results are reported using the AlexNet architecture, pre-trained with ImageNet RGB classification data. For our detection pipeline, we use the recently proposed Fast R-CNN [16] algorithm. We train both the RGB and depth networks each for 40,000 iterations with learning rate 0.001, momentum 0.9, and weight decay 0.0005 using the standard deep learning software package, Caffe [27].

B. RGB-D Detection

We begin by evaluating our algorithm on the NYUD2 test set for the RGB-D detection task [20]. Since we would like to understand the ability of our algorithm to produce an RGB-D detection model when no depth data is available for direct training, we perform hold one category out experiments. We perform 19 experiments where in the i^{th} experiment we remove labeled depth data corresponding to the i^{th} category when training [1] (so the detector has access to RGB data from all 19 categories and depth data from only 18 categories). We then use these detectors to report the AP obtained on the i^{th} category. The performance obtained by our method is reported in Table I under the name 'RGB + aux D'.

We compare against both the Fast R-CNN [16] RGB-only baseline as well as the state-of-the-art RGB-D detection models from Gupta *et. al.*([19] and [20] + Fast R-CNN as described in [21]). Note that the later algorithms require full RGB and depth annotations and as such serve as an upper bound performance for our detection scenario. For reference, we also train our network using full RGB-D training data and report the performance as the oracle for our method (see Table II). This number is expected to be slightly lower than competing state-of-the-art methods since our overall architecture ignores the semantic information learned in the highest layers of the depth network. This is necessary for the held out depth scenario, but is limiting in the full annotation scenario.

Overall, our method achieves 33.8% mAP when averaged across each independent held out category. In comparison RGB only model (but with the same MCG RGB-D proposals) only obtains a mAP of 27.8%. This shows that our mid-level fusion of RGB and depth is able to extract meaningful depth information which can be effectively combined with the RGB information to improve the eventual labeling function.

C. Ablation Study

In this section, we perform an ablation study on the architecture merge layer selection. For this experiment we further split the training set into the standard train/val sets, training with the train set and evaluating on the validation set. Table II reports results on the NYUD2 validation data set for our algorithm while varying the merge point of the RGB and depth networks. We select between the spatially aware pool5 layer and the higher, more semantically meaningful, fully connected layers, fc6, fc7, and fc8 (for oracle only).

We run our algorithm using the same experimental setup of holding out depth training data for one category at a time. For reference, we additionally report the performance of the oracle full depth trained network using each of these merge point selections. We find that merging the RGB and depth networks after fc6 provides the most benefit over using hte RGB-only network. Since the depth network was trained only on the auxiliary 18 object categories, all category specific information which has been stored in the fc7 parameters serves as a distraction when attempting to detect the held out category.

In contrast, the oracle network performs best when merged after fc8, in other words a pure late-fusion approach. This is because the category specific parameters are relevant for all categories we wish to detect and are complementary to the RGB category specific parameters and aid the detection model at test time. Note that this is slightly different than the performance for Gupta *et. al.*[20] + Fast R-CNN reported in Table I. In our experiments the depth network was finetuned from the RGB network already finetuned on NYUD2 RGB images, as opposed to Gupta *et. al.*[20] + Fast R-CNN which was finetuned from ImageNet classification weights.

D. Error Analysis

To investigate how our method uses depth to improve detection, we analyze the false positive errors made by our RGB-D detectors as compared to the baseline RGB-only and oracle fully supervised RGB-D detectors.

We know from Table I that our algorithm has fewer false positives overall than the RGB baseline and has more false positives than the oracle fully supervised RGB-D model. For further insight, we analyze the change in each type of false positives between our method and the baseline and between the oracle and baseline methods (see Figure 3). More precisely, for a given category, i, which has K ground truth instances in the test set, we look at the top K scoring

[1] We do this by removing all bounding box proposals that overlap with the ground truth boxes for category i by any amount, though due to the small dataset size we continue to use regions from the image which do not overlap with the held out category. Note that a held out object may appear within the receptive field of another completely non-overlapping positive object or background box proposal due to the large size of pool5 receptive fields, but there is no supervision for the held-out category.

method	modality	bath tub	bed	book shelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	tele vision	toilet	mean
Girshick *et. al.*(Fast R-CNN) [16]	RGB	7.9	51.2	37.0	1.5	31.3	35.4	9.4	22.4	28.9	19.3	31.0	35.9	24.1	26.4	24.6	39.7	16.6	32.9	53.5	27.8
Our method merge fc6	RGB + aux D	9.7	64.1	37.4	2.1	40.2	44.8	11.9	21.7	39.2	27.8	35.4	46.8	40.2	36.8	27.9	48.4	22.8	35.5	49.0	33.8
Oracle merge fc6	RGB + D	4.7	73.3	45.6	3.6	45.2	54.6	16.8	26.1	47.1	34.9	40.8	49.7	51.7	41.6	39.3	55.7	27.3	48.5	64.4	40.6
Gupta *et. al.* [19]	RGB + D	39.4	73.6	38.4	5.9	50.1	47.3	14.6	24.4	42.9	51.5	36.2	52.1	41.5	42.9	42.6	54.6	25.4	48.6	50.2	41.2
Gupta *et. al.* [20] + Fast R-CNN	RGB + D	37.1	78.3	48.5	3.3	45.3	54.6	21.9	28.5	48.6	41.9	42.5	60.6	49.2	43.7	40.2	62.1	29.2	44.3	63.6	44.4

TABLE I

RGB-D DETECTION (MEAN AP%) ON NYUD2 TEST SET: WE COMPARE OUR PERFORMANCE AGAINST SEVERAL STATE-OF-THE-ART METHODS. ALL METHODS USE THE ALEXNET ARCHITECTURE. TO REPORT PERFORMANCE OF OUR METHOD 'RGB + AUX D' FOR A PARTICULAR CATEGORY c, WE USE RGB DATA FROM ALL 19 CATEGORIES AND DEPTH DATA FROM THE REMAINING 18 'AUXILIARY' CATEGORIES FOR TRAINING. WE SEE OUR METHOD IS ABLE TO IMPROVE PERFORMANCE ON HELD OUT CATEGORIES FROM 27.8% TO 33.8%.

method	modality	merge point	bath tub	bed	book shelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	tele vision	toilet	mean
Baseline [16]	RGB	-	9.2	44.1	11.6	1.4	24.4	25.0	6.8	17.8	15.0	18.6	18.1	42.1	25.3	16.3	19.6	21.0	13.6	35.5	58.5	22.3
Ours	RGB + aux D	pool5	8.4	50.2	3.4	2.1	26.8	25.7	3.6	10.0	23.9	33.5	14.1	38.7	36.1	23.3	20.1	28.8	14.8	31.6	61.6	24.0
Oracle	RGB + D	pool5	11.7	57.7	5.6	2.7	29.9	31.9	4.5	13.0	28.4	42.8	30.3	39.6	39.6	32.5	24.0	33.8	18.3	32.7	63.8	28.6
Ours	RGB + aux D	fc6	8.4	54.0	11.0	1.7	27.5	28.6	6.8	16.8	27.1	30.8	20.3	46.0	40.5	24.0	22.6	30.8	17.3	36.6	64.6	27.1
Oracle	RGB + D	fc6	14.3	66.1	16.9	3.0	36.4	39.3	6.8	20.2	31.9	39.2	31.6	45.1	48.1	32.8	28.6	38.9	22.9	37.7	69.1	33.1
Ours	RGB + aux D	fc7	4.7	54.3	6.3	1.1	26.4	26.4	5.7	9.3	27.6	21.9	15.2	44.2	35.6	15.6	8.8	28.8	16.3	35.8	54.0	23.1
Oracle	RGB + D	fc7	14.9	67.0	19.7	3.0	37.5	38.9	8.2	18.3	31.9	34.0	35.0	45.4	50.3	36.3	30.9	41.4	22.8	37.5	71.2	33.9
Oracle	RGB + D	fc8	15.4	70.6	21.6	3.7	37.4	38.2	8.8	17.4	31.1	34.4	36.7	43.6	50.7	37.5	30.2	40.4	22.9	38.1	71.5	34.2

TABLE II

RGB-D DETECTION (MAP%) ON NYUD2 VAL SET: WE COMPARE VARIOUS ARCHITECTURE MERGE POINTS. ALL METHODS USE THE ALEXNET ARCHITECTURE AND HOLD DEPTH TRAINING DATA OUT FOR THE CATEGORY BEING STUDIED. WE FIND THAT MERGING AFTER FC6 PERFORMS THE BEST ON THIS DATASET FOR THE MISSING DATA SETTING. HOWEVER, WHEN ALL DEPTH TRAINING DATA IS AVAILABLE, LATE FUSING AT THE SCORES IS THE BEST OPTION.

regions across the test set from the category i detector from the baseline RGB-only model, our model, and the oracle RGB-D model. For each model we compute the percent of the top K detections which correspond to each type of false positives. We then plot the difference in this percentage between the baseline and our method and the baseline and the oracle. For ease of viewing, categories are sorted per false positive type from least improvement of our method to most improvement by our method.

By studying these changes in the false positives, some interesting trends emerge. For instance, we find that our approach provides a relatively consistent improvement in localization and confusion with other categories (most bars in the *top row* are greater than or close to zero). In contrast, our method improves only 11/19 of the categories in confusion with background and hinders performance for the other 8/19 categories. We see that the oracle method provides improvement in confusion with background for almost all categories, which indicates there is potential to further improve these types of errors when RGB-D training data is available for the category of interest.

One interesting category is *television*, which has over a 15% reduction in the confusion with background when using our algorithm over the RGB baseline, but simultaneously has almost a 15% increase in the confusion with other category false positives. This is likely due to the fact that *monitor* is another category available and since during depth training the held out category *television* is not seen at the same time as the known category, *monitor*, this makes it harder for our algorithm to disambiguate the two categories at test time. This issue is mitigated with full supervision training of the depth net.

Finally, we show some qualitative examples of the improvements made by our approach. We pick the two categories where our method improves the most and least over the baseline. Figure 4 shows random images which contain bed and night-stand (categories where we improve the most 12.9% and 16.1%) where the top scoring detection is a true positive for our method and false positive for the baseline. Similarly in Figure 5 we show random images containing toilets and doors (categories where we improve the least -4.5% and -0.7%) where the highest scoring detection is a true positive for the baseline while it is false positive for our method.

In Figure 4, we very clearly see the effects of our method improving localization errors as well as fixing confusion

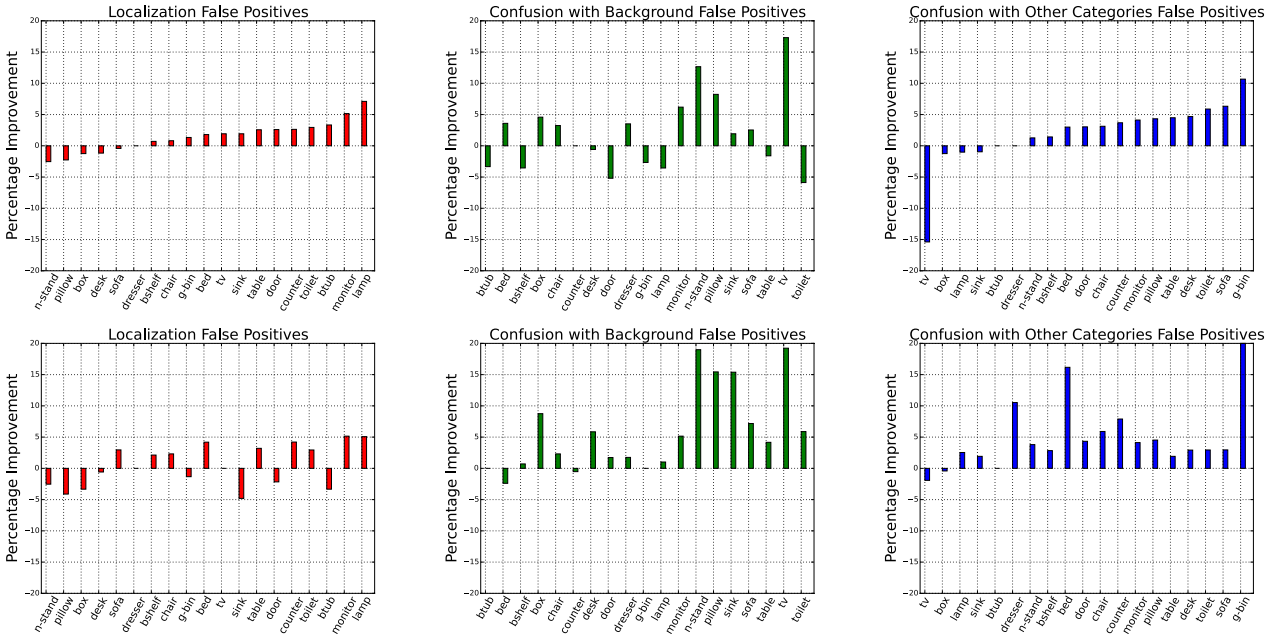

Fig. 3. We study the change in the type of false positives between baseline and our method (*top row*) and the the change in the type of false positives between the baseline and the oracle for our method (*bottom row*). We show here false positives due to localization errors (red - left), confusion with background (green - center), and confusion with other categories (blue - right).

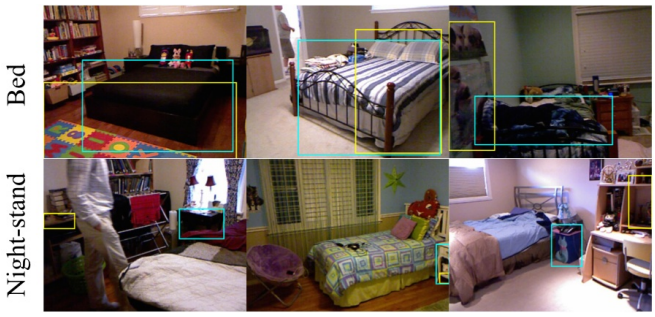

Fig. 4. Example detections on the NYUD2 test set where the top detection from our method for the specified category is correct while the top detection from the RGB only baseline is incorrect. Cyan boxes are from our method and yellow boxes are from the RGB baseline.

Fig. 5. Example detections on the NYUD2 test set where the top detection from the RGB only baseline for the specified category is correct while the top detection from our method is incorrect. Cyan boxes are from our method and yellow boxes are from the RGB baseline.

with other categories. Similarly, Figure 5, provides examples where we begin to confuse with non-objects (background) for the toilet and door categories. With the exception of one of the toilet examples (middle) which is simply a result of the baseline region being just over threshold for overlap with ground truth to be considered a true positive, while our method's top scoring example was just under the threshold.

E. Large Scale RGB-D Detection

One of the main motivations behind our work is to enable enhanced RGB-D detection of a large number of objects with no depth training data, for applications such as robotics. We demonstrate the potential impact of our work by using our algorithm to extend the released 7.6k RGB detector [24] into an RGB-D detector, and show qualitative results in Figure 6. The LSDA [24] model was available only for RGB detection along with an RGB region proposal method (selective search [49]). We show results for the model from [24] in the *left* column. Next, we use the network parameters from the model from [24] along with RGB-D MCG proposals, as used throughout our method – the results are displayed in the *center* column. Finally, we produce a joint RGB-D network through our method of mid-level representation fusion and show results for our algorithm in the *right* column.[2] We show results on images taken from two scenes in the Cornell activity dataset [45], which contains categories not available during training on NYUD2 data, such as person.

After changing the region proposal mechanism to incorporate depth information, we see significant improvement in

[2]Note that these results were obtained using the publicly released LSDA R-CNN detector [24] and not the Fast R-CNN detector that is used for the rest of the experiments. We expect similar results with the Fast R-CNN based detector.

Fig. 6. We use our algorithm to transform the publicly available 7.6k class RGB detector [24] into an RGB-D detector. We show here detection results for all 7.6k categories on example RGB-D images taken from two scenes in the Cornell activity dataset [45]. We present top detections from the original RGB CNN with RGB selective search region proposals (*left*), detections when using RGB-D MCG proposals (*middle*), and detections after our proposed adaptation (*right*). Blue boxes are detections of the 200 ILSVRC categories, while the red boxes are detections of the 7.4k categories corresponding to leaf nodes in the ImageNet database. Our algorithm not only provides better localization, but even enables extra categories to be detected.

object localization. Upon using our algorithm to transform the RGB network into an RGB-D network, we see that false positives are reduced and new objects are recognized.

This qualitative result is highly encouraging as it demonstrates that our algorithm incorporates category invariant depth information that is generic enough to be useful with a detector that was trained on separate tasks and in a different data source. For example, people, shower stalls, and credenzas never appear in NYUD2 training annotations, where we train our depth model. However, we are able to learn to effectively combine the generic depth and RGB processing of the lower layers and use the modified intermediate representation as additional information for the category specific classification layer. This model was able to be produced without further RGB training, meaning that our pre-trained RGB detector could immediately be adapted to utilize depth information at test time. In the future we plan to conduct a more quantitative study of this results.

V. Conclusion

We have presented an algorithm that can transform an RGB object detector into a RGB-D detector which can use depth data at test time to improve performance. Our multi-modal CNN architecture combines mid-level RGB and depth representations to incorporate both modalities into the final object class prediction. This mid-level fusion enables us to train RGB-D detectors without needing complete RGB-D data, unlike most conventional CNN based RGB-D object detection algorithms.

We present experiments showing that our approach provides a 21% relative improvement in performance over just using an RGB detector for categories without no depth data available at training time. We provide insight on how our system helps improve object detection compared to RGB-only detection. Finally, we use our algorithm to adapt the 7.6k category detectors from [24] into a multi-modal RGB-D version, and show qualitative results with this large scale depth detector.

Experiments thus far have been presented using the two stage region proposals and CNN-based feature computation per region, as introduced in R-CNN [15] and Fast R-CNN [16]. Our final goal is to provide a system which can be practically used in a robotics setting. In the future we will work towards making our detectors faster possibly with the use of end-to-end CNN object detection systems like Faster R-CNN [38] and more accurate with use of better CNNs for depth images [21].

References

[1] K. Ali and K. Saenko. Confidence-rated multiple instance boosting for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Proc. NIPS*, pages 561–568, 2002.

[3] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *In: CVPR (2014)*.

[4] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.

[5] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[6] S. Chopra, S. Balakrishnan, and R. Gopalan. DLID: Deep learning for domain adaptation by interpolating between domains. In *ICML Workshop on Challenges in Representation Learning*, 2013.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[8] C. Mario Chrisoudias, Raquel Urtasun, Mathieu Salzmann, and Trevor Darrell. Learning to recognize objects from unseen modalities. In *ECCV*, 2010.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Piotr Dollár and C. Lawrence Zitnick. Fast edge detection using structured forests. *CoRR*, abs/1406.5549, 2014.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proc. ICML*, 2014.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[13] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Tran. PAMI*, 32(9):1627–1645, 2010.

[14] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proc. ICCV*, 2013.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proc. CVPR*, 2014.

[16] Ross Girshick. Fast R-CNN. *arXiv preprint arXiv:1504.08083*, 2015.

[17] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012.

[18] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 564–571. IEEE, 2013.

[19] Saurabh Gupta, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014*, pages 345–360. Springer, 2014.

[21] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. *CoRR*, abs/1507.00448v1, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *In Proc. ECCV*, 2014.

[23] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.

[24] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. arXiv:1407.5035, 2014.

[25] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *BMVC*, 2014.

[26] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3D object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*. 2013.

[27] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[28] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *Computer Vision–ECCV 2014*, pages 725–739. Springer, 2014.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[30] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

[31] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *IEEE International Conference on on Robotics and Automation*, 2014.

[32] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[33] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3D object detection with RGBD cameras. In *ICCV*, 2013.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Pitro Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *arXiv:1405.0312 [cs.CV]*, 2014.

[36] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[37] Wanli Ouyang, Ping Luo, Xingyu Zeng, Shi Qiu, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Yuanjun Xiong, Chen Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv:1409.3505*, 2014.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015.

[39] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.

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

[41] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[42] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *Computer Vision–ECCV 2014*, pages 634–651. Springer, 2014.

[43] Byung soo Kim, Shili Xu, and Silvio Savarese. Accurate localization of 3D objects from RGB-D data using segmentation hypotheses. In *CVPR*, 2013.

[44] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[45] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgbd images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849. IEEE, 2012.

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[47] Jie Tang, Stephen Miller, Arjun Singh, and Pieter Abbeel. A textured object recognition pipeline for color and depth image data. 2012.

[48] Shuai Tang, Xiaoyu Wang, Xutao Lv, Tony X Han, James Keller, Zhihai He, Marjorie Skubic, and Shihong Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *ACCV*, 2012.

[49] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[50] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision (ECCV)*, 2014.

[51] Edmund Shanming Ye. Object detection in rgb-d indoor scenes. Master's thesis, EECS Department, University of California, Berkeley, Jan 2013.

[52] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014.