

A deep representation for depth images from synthetic data

Fabio Maria Carlucci¹ and Paolo Russo¹ and Barbara Caputo¹

Abstract—Convolutional Neural Networks (CNNs) trained on large scale RGB databases have become the secret sauce in the majority of recent approaches for object categorization from RGB-D data. Thanks to colorization techniques, these methods exploit the filters learned from 2D images to extract meaningful representations in 2.5D. Still, the perceptual signature of these two kind of images is very different, with the first usually strongly characterized by textures, and the second mostly by silhouettes of objects. Ideally, one would like to have two CNNs, one for RGB and one for depth, each trained on a suitable data collection, able to capture the perceptual properties of each channel for the task at hand. This has not been possible so far, due to the lack of a suitable depth database. This paper addresses this issue, proposing to opt for synthetically generated images rather than collecting by hand a 2.5D large scale database. While being clearly a proxy for real data, synthetic images allow to trade quality for quantity, making it possible to generate a virtually infinite amount of data. We show that the filters learned from such data collection, using the very same architecture typically used on visual data, learns very different filters, resulting in depth features (a) able to better characterize the different facets of depth images, and (b) complementary with respect to those derived from CNNs pre-trained on 2D datasets. Experiments on two publicly available databases show the power of our approach.

I. INTRODUCTION

Deep learning has changed the research landscape in visual object recognition over the last few years. Since their spectacular success in recognizing 1,000 object categories [1], convolutional neural networks have become the new off the shelf state of the art in visual classification. Since then, the robot vision community has also attempted to take advantage of the deep learning trend, as the ability of robots to understand what they see reliably is critical for their deployment in the wild. A critical issue when trying to transfer results from computer to robot vision is that robot perception is tightly coupled with robot action. Hence, pure RGB visual recognition is not enough.

The heavy use of 2.5D depth sensors on robot platforms has generated a lively research activity on 2.5D object recognition from depth maps [2], [3], [4]. Here a strong emerging trend is that of using Convolutional Neural Networks (CNNs) pre-trained over ImageNet [5] by colorizing the depth channel [6]. The approach has proved successful, especially when coupled with fine tuning [7] and/or spatial pooling strategies

[8], [9], [10] (for a review of recent work we refer to section II). These results suggest that the filters learned by CNNs from ImageNet are able to capture information also from depth images, regardless of their perceptual difference.

Is this the best we can do? What if one would train from scratch a CNN over a very large scale 2.5D object categorization database, wouldn't the filters learned be more suitable for object recognition from depth images? RGB images are perceptually very rich, with generally a strong presence of textured patterns, especially in ImageNet. Features learned from RGB data are most likely focusing on those aspects, while depth images contain more information about the shape and the silhouette of objects. Unfortunately, as of today a 2.5D object categorization database large enough to train a CNN on it does not exist. A likely reason for this is that gathering such data collection is a daunting challenge: capturing the same variability of ImageNet over the same number of object categories would require the coordination of very many laboratories, over an extended period of time.

In this paper we follow an alternative route. Rather than acquiring a 2.5D object categorization database, we propose to use synthetic data as a proxy for training a deep learning architecture specialized in learning depth specific features. To this end, we construct the VANDAL database, a collection of 4.1 million depth images from more than 9,000 objects, belonging to 319 categories. The depth images are generated starting from 3D CAD models, downloaded from the Web, through a protocol developed to extract the maximum information from the models. VANDAL is used as input to train from scratch a deep learning architecture, obtaining a pre-trained model able to act as a depth specific feature extractor. Visualizations of the filters learned by the first layer of the architecture show that the filter we obtain are indeed very different from those learned from ImageNet with the very same convolutional neural network (figure 1). As such, they are able to capture different facets of the perceptual information available from real depth images, more suitable for the recognition task in that domain. We call our pre-trained architecture DepthNet.

Experimental results on two publicly available databases confirm this: when using only depth, our DepthNet features achieve better performance compared to previous methods based on a CNN pre-trained over ImageNet, without using fine tuning or spatial pooling. The combination of the DepthNet features with the descriptors obtained from the CNN pre-trained over ImageNet, on both depth and RGB images, leads to strong results on the Washington database [11], and to results competitive with fine-tuning and/or sophisticated spatial pooling approaches on the JHUIT database [12]. To

*This work was supported by the projects ALOOF CHIS-ERA (F.M.C.) and ERC RoboExNovo (B.C.). We thank S. Baharlou, A. Gigli, F. Giordaniello, M. Graziani and M. Lampacrescia for their help in creating the VANDAL database.

¹F. M. Carlucci, P. Russo and B. Caputo are with the VANDAL Laboratory, Department of Computer, Management and Control Engineering (DIAG), Sapienza Rome University, Rome, Italy fabiom.carlucci@dis.uniroma1.it

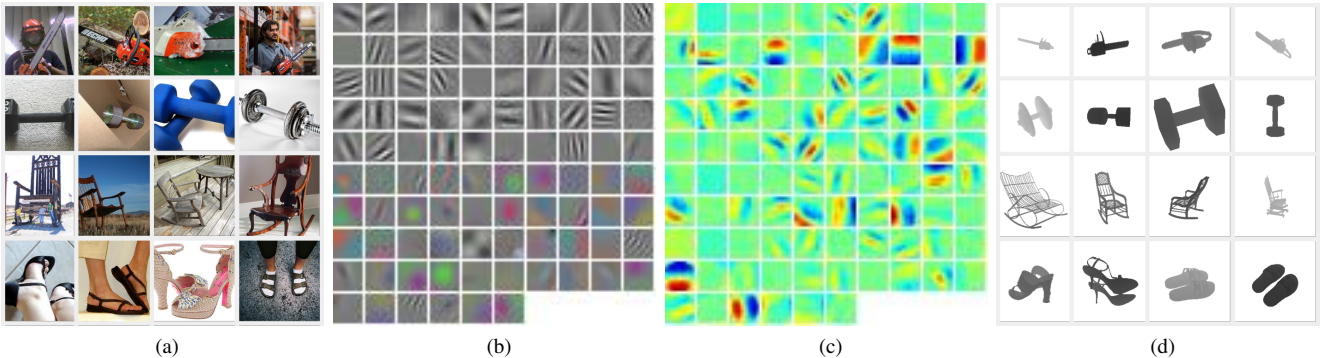


Fig. 1: Sample images for the classes chainsaw, dumbbell, rocker chair and sandal from ImageNet (a) and VANDAL (d). We show the corresponding filters learned by the very same CNN architecture respectively in (b) and (c). We see that even though the architecture is the same, using 2D rather than 2.5D images for training leads to learning quite different filters.

the best of our knowledge, this is the first work that uses synthetically generated depth data to train a depth-specific convolutional neural network. Upon acceptance of the paper, all the VANDAL data, the protocol and the software for generating new depth images, as well as the pre-trained DepthNet, will be made publicly available.

The rest of the paper is organized as follows. After a review of the recent literature (section II), we introduce the VANDAL database, describing its generation protocol and showcasing the obtained depth images (section III). Section IV describes the deep architecture used and section V reports our experimental findings. The paper concludes with a summary and a discussion on future research.

II. RELATED WORKS

Object recognition from RGB-D data traditionally relied on hand-crafted features such as SIFT [13] and spin images [2], combined together through vector quantization in a Bag-of-Words encoding [2]. This heuristic approach has been surpassed by end-to-end feature learning architectures, able to define suitable features in a data-driven fashion [14], [3], [15]. All these methods have been designed to cope with a limited amount of training data (of the order of $10^3 - 10^4$ depth images), thus they are able to only partially exploit the generalization abilities of deep learning as feature extractors experienced in the computer vision community [1], [16], where databases of 10^6 RGB images like ImageNet [5] or Places [17] are available.

An alternative route is that of re-using deep learning architectures trained on ImageNet through pre-defined encoding [18] or colorization. Since the work of [6] re-defined the state of the art in the field, this last approach has been actively and successfully investigated. Eitel et al [7] proposed a parallel CNN architecture, one for the depth channel and one for the RGB one, combined together in the final layers through a late fusion scheme. Some approaches coupled non linear learning methods with various forms of spatial encodings [10], [9], [4], [12]. Hasan et al [8] pushed further this multi-modal approach, proposing an architecture merging together RGB,

depth and 3D point cloud information. Another notable feature is the encoding of an implicit multi scale representation through a rich coarse-to-fine feature extraction approach.

All these works build on top of CNNs pre-trained over ImageNet, for all modal channels. Thus, the very same filters are used to extract features from all of them. As empirically successful as this might be, it is a questionable strategy, as RGB and depth images are perceptually very different, and as such they would benefit from approaches able to learn data-specific features (figure 1). Our method matches this challenge, learning RGB features from RGB data and depth features from synthetically generated data, within a deep learning framework. The use of realistic synthetic data in conjunction with deep learning architectures is a promising emerging trend [19], [20], [21]. We are not aware of previous work attempting to use synthetic data to learn depth representations, with or without deep learning techniques.

III. THE VANDAL DATABASE

In this section we present VANDAL and the protocol followed for its creation. With 4,106,340 synthetic images, it is the largest existing depth database for object recognition. Section III-A describes the criteria used to select the object categories composing the database and the protocol followed to obtain the 3D CAD models from Web resources. Section III-B illustrates the procedure used to generate depth images from the 3D CAD models.



Fig. 2: Sample morphs (center, right) generated from an instance model for the category coffee cup (left).

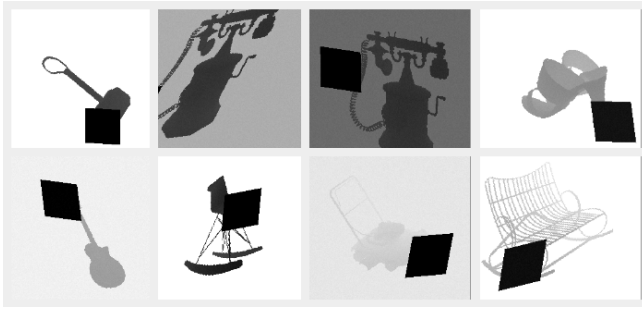


Fig. 5: Data augmentation samples from various classes (hammer, phone, sandal, guitar, rocker, lawn mower, bench). Note that the contrast/brightness variations and noise are hard to visualize on small thumbnails.

model, obtaining a total of 4.1 million images. Preliminary experiments showed that increasing the sampling rate in the configuration space did lead to growing percentages of nearly identical images.

The rendered depth images consist of objects always centered on a white background. This is done on purpose, as it allows us the maximum freedom to perform various types of data augmentation at training time, as it is standard practice when training convolutional neural networks. This is here even more relevant than usual, as synthetically generated data are intrinsically perceptually less informative compared to real data. The data augmentation methods we used are: image cropping, occlusion (1/4 of the image is randomly occluded to simulate gaps in the sensor scan), contrast/brightness variations, in depth views corresponding to scaling the Z axis and shifting the objects along it, background substitution (substituting the white background with one randomly chosen farther away than the object’s center of mass), random uniform noise (as in film grain), and image shearing (a slanting transform). Figure 5 shows some examples of data augmentation images obtained with this protocol.

IV. LEARNING DEEP DEPTH FILTERS

Once the VANDAL database has been generated, it is possible to use it to train any kind of convolutional deep architecture. In order to allow for a fair comparison with previous work, we opted for CaffeNet, a slight variation of AlexNet [1]. Although more modern networks have been proposed in the last years [22], [23], [24], it still represents the most popular choice among practitioners, and the most used in robot vision³. Its well known architecture consists of 5 convolutional layers, interwoven with pooling, normalization and relu layers, plus three fully connected layers. CaffeNet differs from AlexNet in the pooling, which is done there before normalization. It usually performs slightly better and has thus gained wide popularity.

³Preliminary experiments using the VGG, Inception and Wide Residual networks on the VANDAL database did not give stable results and need further investigation.

Although the standard choice in robot vision is using the output of the seventh activation layer as feature descriptors, several studies in the vision community show that lower layers, like the sixth and the fifth, tend to have higher generalization properties [25]. We followed this trend, and opted for the fifth layer (by vectorization) as deep depth feature descriptor (an ablation study supporting this choice is reported in section V). We name in the following as **DepthNet** the CaffeNet architecture trained on VANDAL using as output feature the fifth layer, and **Caffe-ImageNet** the same architecture trained over ImageNet.

Once DepthNet has been trained, it can be used as any depth feature descriptor, alone or in conjunction with Caffe-ImageNet for classification of RGB images. We explore this last option, proposing a system for RGB-D object categorization that combines the two feature representations through a multi kernel learning classifier [26]. Figure 6 gives an overview of the overall RGB-D classification system. Note that DepthNet can be combined with any other RGB and/or 3D point cloud descriptor, and that the integration of the modal representations can be achieved through any other cue integration approach. This underlines the versatility of DepthNet, as opposed to recent work where the depth component was tightly integrated within the proposed overall framework, and as such unusable outside of it [7], [8], [4], [12].

V. EXPERIMENTS

We assessed the DepthNet, as well as the associated RGB-D framework of figure 6, on two publicly available databases. Section V-A describes our experimental setup and the databases used in our experiments. Section V-B reports a set of experiments assessing the performance of DepthNet on depth images, compared to Caffe-ImageNet, while in section V-C we assess the performance of the whole RGB-D framework with respect to previous approaches.

A. Experimental setup

We conducted experiments on the Washington RGB-D [11] and the JHUIT-50 [12] object datasets. The first consists of 41,877 RGB-D images organized into 300 instances divided in 51 classes. Each object instance was positioned on a turntable and captured from three different viewpoints while rotating. Since two consecutive views are extremely similar, only 1 frame out of 5 is used for evaluation purposes. We performed experiments on the object categorization setting, where we followed the evaluation protocol defined in [11]. The second is a challenging recent dataset that focuses on the problem of fine-grained recognition. It contains 50 object instances, often very similar with each other (e.g. 9 different kinds of screwdrivers). As such, it presents different classification challenges compared to the Washington database.

All experiments, as well as the training of DepthNet, were done using the publicly available Caffe framework [27], together with NVIDIA Deep Learning GPU Training System (DIGITS). As described above, we obtained DepthNet by training a CaffeNet over the VANDAL database. The network

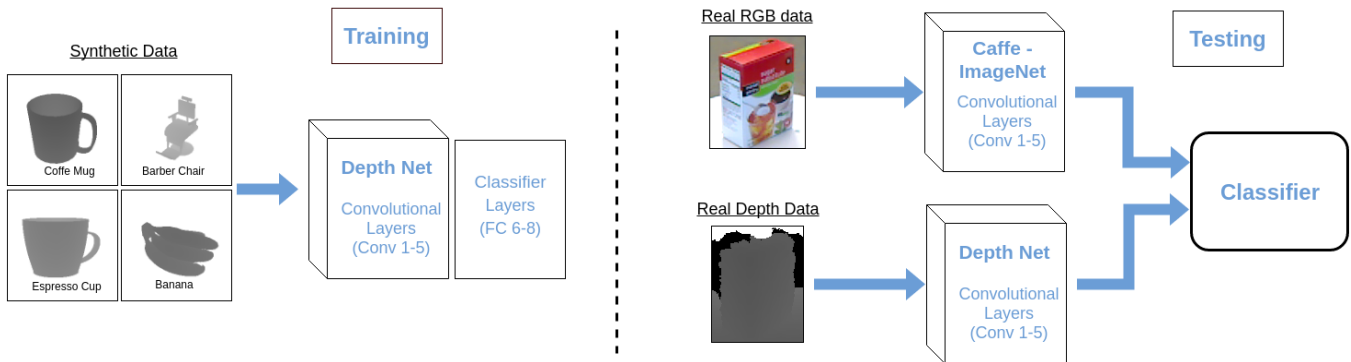


Fig. 6: DepthNet and our associated RGB-D object classification framework. During training, we learn depth filters from the VANDAL synthetic data (left). During test (right), real RGB and depth data is processed by two distinct CNNs, each specialized over the corresponding modality. The features, derived from the activations of the fifth convolutional layer, are then fed into a cue integration classifier.

was trained using Stochastic Gradient Descent for 50 epochs. Learning rate started at 0.01 and gamma at 0.5 (halving the learning rate at each step). We used a variable step down policy, where the first step took 25 epochs, the next 25/2, the third 25/4 epochs and so on. These parameters were chosen to make sure that the test loss on the VANDAL test data had stabilized at each learning rate. Weight decay and momentum were left at their standard values of 0.0005 and 0.9.

To assess the quality of the DepthNet features we performed three set of experiments:

- 1) *Object classification using depth only*: features were extracted with DepthNet and a linear SVM⁴ was trained on it. We also examined how the performance varies when extracting from different layers of the network, comparing against a Caffe-ImageNet used for depth classification, as in [6].
- 2) *Object classification using RGB + Depth*: in this setting we combined our depth features with those extracted from the RGB images using Caffe-ImageNet. While [7] train a fusion network to do this, we simply use an off the shelf Multi Kernel Learning (MKL) classifier [26].

For all experiments we used the training/testing splits originally proposed for each given dataset. For linear SVM, we set C by cross validation. When using MKL, we left the default values of 100 iterations for online and 300 for batch and set p and C by cross validation.

Previous works using Caffe-ImageNet as feature extractor for depth, apply some kind of input preprocessing [6], [7], [8]. While we do compare against the published baselines, we also found that by simply normalizing each image (min to 0 and max to 255), one achieves very competitive results. Also, since our DepthNet is trained on depth data, it does not need any type of preprocessing over the depth images, obtaining strong results over raw data. Because of this, in all experiments reported in the following we only consider raw depth images and normalized depth images.

⁴Liblinear: <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>

B. Assessing the performance of the DepthNet architecture

We present here an ablation study, aiming at understanding the impact of choosing features from the last fully convolutional layer as opposed to the more popular last fully connected layer, and of using normalized depth images instead of raw data. By comparing our results with those obtained by Caffe-ImageNet, we also aim at illustrating up to which point the features learned from VANDAL are different from those derived from ImageNet.

Figure 7 shows results obtained on the Washington database, with normalized and raw depth data, using as features the activations of the fifth pooling layer (pool5), of the sixth fully connected layer (FC6), and of the seventh fully connected layer (FC7). Note that this last set of activations is the standard choice in the literature. We see that for all settings, pool5 achieves the best performance, followed by FC6 and FC7. This seems to confirm recent findings on RGB data [25], indicating that pool5 activations offer stronger generalization capabilities when used as features, compared to the more popular FC7. The best performance is obtained by DepthNet, pool5 activations over raw depth data, with a 83.8% accuracy. DepthNet achieves also better results compared to Caffe-ImageNet over normalized data. To get a better feeling of how performance varies when

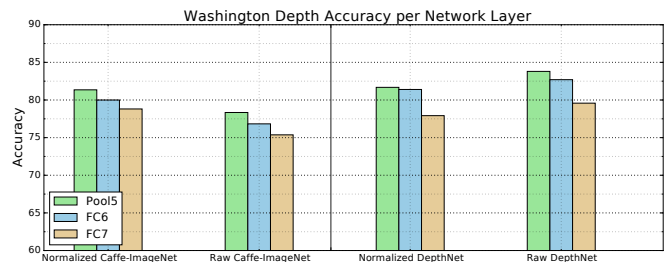
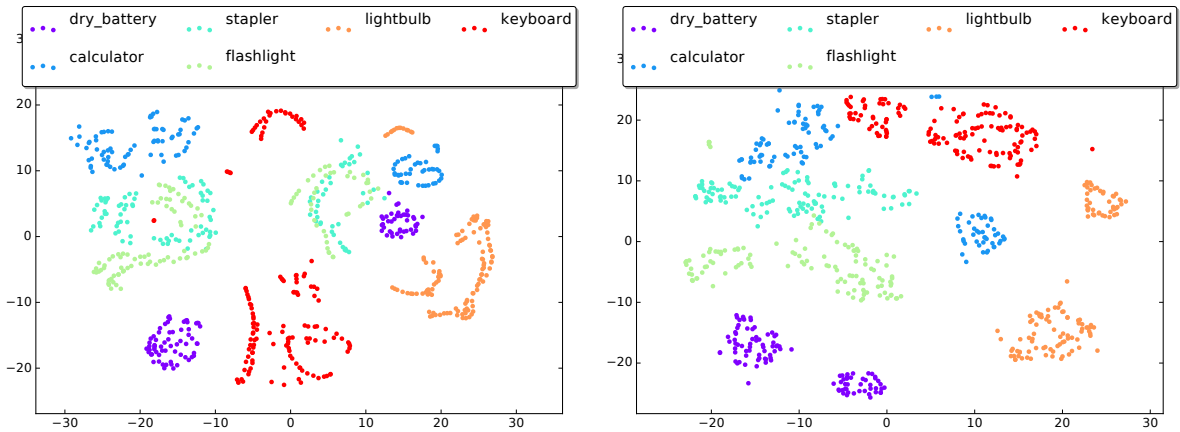
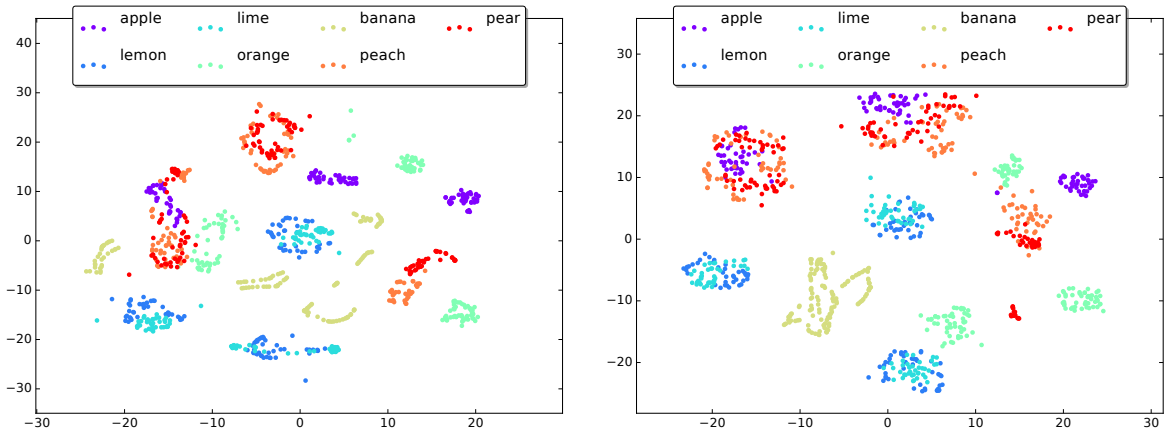


Fig. 7: Accuracy obtained by DepthNet and Caffe-ImageNet over the Washington database, using as features pool5, FC6 and FC7. Results are reported for raw and normalized depth images.



(a) Device classes as seen by Caffe-ImageNet (left) and DepthNet (right)



(b) Fruit classes as seen by Caffe-ImageNet (left) and DepthNet (right)

Fig. 9: t-SNE visualizations for the categories device (top) and fruit (bottom).

Method:	RGB	Depth Mapping	Depth Raw	RGB-D
DepthNet RGB-D Framework	88.49 ± 1.8	81.68 ± 2.2	83.8 ± 2.0	92.25 ± 1.3
Caffe-ImageNet Pool5	88.49 ± 1.8	81.11 ± 2	78.35 ± 2.5	90.79 ± 1.2
Caffe-ImageNet FC7 finetuning[7]	84.1 ± 2.7	83.8 ± 2.7	—	91.3 ± 1.4
Caffe-ImageNet FC7[6]	83.1 ± 2.0	—	—	89.4 ± 1.3
CNN only[4]	82.7 ± 1.2	78.1 ± 1.3	—	87.5 ± 1.1
CNN + FisherKernel + SPM[4]	86.8 ± 2.2	85.8 ± 2.3	—	91.2 ± 1.5
CNN + Hypercube Pyramid + EM[8]	87.6 ± 2.2	85.0 ± 2.1	—	91.4 ± 1.4
CNN-SPM-RNN+CT[10]	85.2 ± 1.2	83.6 ± 2.3	—	90.7 ± 1.1
CNN-RNN+CT[9]	81.8 ± 1.9	77.7 ± 1.4	—	87.2 ± 1.1
CNN-RNN[29]	80.8 ± 4.2	78.9 ± 3.8	—	86.8 ± 3.3

TABLE I: Comparison of our DepthNet framework with previous work on the Washington database. With *depth mapping* we mean all types of depth preprocessing used in the literature.

very important for this kind of tasks [30], [31]. We are inclined to attribute to this the superior performance of [12]; future work incorporating spatial pooling in our framework, as well as further experiments on the object identification task in the Washington database and on other RGB-D data collections will explore this issue.

VI. CONCLUSIONS

In this paper we focused on object classification from depth images using convolutional neural networks. We argued that, as effective as the filters learned from ImageNet

are, the perceptual features of 2.5D images are different, and that it would be desirable to have deep architectures able to capture them. To this purpose, we created VANDAL, the first depth image database synthetically generated, and we showed experimentally that the features derived from such data, using the very same CaffeNet architecture widely used over ImageNet, are stronger while at the same time complementary to them. This result, together with the public release of the database, the trained architecture and the protocol for generating new depth synthetic images, is the

Method:	RGB	Depth Mapp.	Depth Raw	RGB-D
DepthNet Pool5	—	54.37	55.0	90.3
Caffe-ImageNet Pool5	88.05	53.6	38.9	89.6
Caffe-ImageNet FC7[6]	82.08	47.87	26.11	83.6
CSHOT + Color pooling + MultiScale Filters[12]	—	—	—	91.2
HMP[12]	81.4	41.1	—	74.6

TABLE II: Comparison of our DepthNet framework with previous work on the JHUIT database. As only one split is defined, we do not report std.

contribution of this paper.

We see this work as the very beginning of a long research thread. By its very nature, DepthNet could be plugged into all previous work using CNNs pre-trained over ImageNet for extracting depth features. It might substitute that module, or it might complement it; the open issue is when this will prove beneficial in terms of spatial pooling approaches, learning methods and classification problems. A second issue we plan to investigate is the impact of the deep architecture over the filters learned from VANDAL. While in this work we chose on purpose to not deviate from CaffeNet, it is not clear that this architecture, which was heavily optimized over ImageNet, is able to exploit at best our synthetic depth database. While preliminary investigations with existing architectures have not been satisfactory, we believe that architecture surgery might lead to better results. Finally, we believe that the possibility to use synthetic data as a proxy for real images opens up a wide array of possibilities: for instance, given prior knowledge about the classification task of interest, would it be possible to generate on the fly a task specific synthetic database, containing the object categories of interest under very similar imaging conditions, and train and end-to-end deep network on it? How would performance change compared to the use of network activations as done today? Future work will focus on these issues.

REFERENCES

- [1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
- [2] K. Lai, L. Bo, X. Ren, and D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, in Proc. ICRA, 2011, pp. 18171824.
- [3] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Ng, Convolutional-recursive deep learning for 3D object classification, in Proc. NIPS, 2012, pp. 665673.
- [4] Cheng, Yanhua, et al. "Convolutional fisher kernels for rgb-d object recognition." *3D Vision (3DV)*, 2015 International Conference on. IEEE, 2015.
- [5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2014). ImageNet large scale visual recognition challenge. arXiv: 1409 . 0575.
- [6] Schwarz, Max, Hannes Schulz, and Sven Behnke. "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features." 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015.
- [7] Eitel, Andreas, et al. "Multimodal deep learning for robust rgb-d object recognition." *Intelligent Robots and Systems (IROS)*, 2015 IEEE/RSJ International Conference on. IEEE, 2015.
- [8] H. F. M. Zaki, F. Shafait, A. Mian, Convolutional hypercube pyramid for accurate RGB-D object category and instance recognition in Proc.International Conference on Robots and Automation (ICRA), 2016.

- [9] Y. Cheng, X. Zhao, K. Huang, and T. Tan. Semisupervised learning for rgb-d object recognition. In ICPR, 2014.
- [10] Y. Cheng, X. Zhao, K. Huang, and T. Tan. Semisupervised learning and feature evaluation for rgb-d object recognition. *Computer Vision and Image Understanding*, 2015.
- [11] Lai, Kevin, et al. "A large-scale hierarchical multi-view rgb-d object dataset." *Robotics and Automation (ICRA)*, 2011 IEEE International Conference on. IEEE, 2011.
- [12] Li, Chi, Austin Reiter, and Gregory D. Hager. "Beyond spatial pooling: fine-grained representation learning in multiple domains." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [13] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol. 60, no. 2, pp.91110, 2004.
- [14] M. Blum, J. T. Springenberg, J. Wulffing, and M. Riedmiller, A learned feature descriptor for object recognition in RGB-D data, in Proc. ICRA, 2012, pp. 12981303.
- [15] U. Asif, M. Bennamoun, and F. Sohel, Efficient RGB-D object categorization using cascaded ensembles of randomized decision trees, in Proc. ICRA, 2015, pp. 12951302.
- [16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in Proc. Computer Vision and Pattern Recognition Workshops (CVPRW), 2014, pp. 512519.
- [17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.
- [18] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in Proc. ECCV, 2014, pp. 345360.
- [19] J. Papon, M. Schoeler. Semantic pose using deep networks trained on synthetic RGB-D. Proc. International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [20] D. Maturana, S. Scherer. VoxNet: a 3D convolutional neural network for real-time object recognition. Proc International Conference on Robots and Systems (IROS), 2015.
- [21] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling. Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)
- [22] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
- [23] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [24] He, Kaiming, et al. "Deep residual learning for image recognition." arXiv preprint arXiv:1512.03385 (2015).
- [25] Zheng, Liang, et al. "Good Practice in CNN Feature Transfer." arXiv preprint arXiv:1604.00133 (2016).
- [26] Orabona, Francesco, Luo Jie, and Barbara Caputo. "Online-batch strongly convex multi kernel learning." *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on. IEEE, 2010.
- [27] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.
- [28] L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research*, vol. 9, pp. 25792605, 2008.
- [29] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Ng, Convolutional-recursive deep learning for 3d object classification. In NIPS, 2012
- [30] Zhang, Ning, Ryan Farrell, and Trevor Darrell. "Pose pooling kernels for sub-category recognition." *Computer Vision and Pattern Recognition (CVPR)*, 2012 IEEE Conference on. IEEE, 2012.
- [31] Angelova, Anelia, and Philip M. Long. "Benchmarking large-scale fine-grained categorization." *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014.