Learn transferable features for semantic image segmentation in the presence of label noise

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

July 21, 2017

Abstract

The ABSTRACT HERE

1 Introduction

Why noisy labels?:This paragraph should discuss the difficulties for collecting perfect segmentation annotations on a large scale

The recent success of deep neural networks benefits from the availability of large-scale supervised datasets such as ILSVRC[20]. However, collecting a dataset for semantic image segmentation on a large scale can be expensive and time-consuming, especially when error-free labels are required. Enormous efforts have been made to create the "gold standard" annotations for the current benchmark segmentation datasets[5, 16, 12]. [1] These datasets allow the deep neural network models for semantic segmentation [13, 25] to develop assuming the perfect segmentation ground truth exists, where perfect means that all the instances are annotated, no non-target object is misannotated, and no instance is misclassified. However, it is natural for human beings to make mistakes due to the lack of expertise, the inherent ambiguity of tasks or unconscious bias. Strong supervision is required to correct the mistakes, including double-checking the annotations over and over again and ensembling opinions from multiple annotators. In some cases, for example in the medical imaging, the "gold standard" itself can be ambiguous and cause disagreement among different annotators. Also, there are some freely available labels, which may or may not be accurate, for specific problems. For example, one can use digital maps, like OpenStreetMap, to annotate aerial images, but such segmentation annotations constructed from maps suffer from the incomplete annotation as well as registration problems.[15] This motivates us to explore how to learn from these noisy annotations.

Why representation?: This paragraph should discuss the idea of noise-robust representation

Deep convolutional neural networks (DCNN) based models often contain two principle components: a stack of convolutional layers to extract hierarchical features and a few task-specific layers to fit the training objectives. Both parts are important to achieve good performance because they are trained jointly in an end-to-end manner via back propagation. The convolutional features were proved "transferable" to a new dataset[24] or even to a different task[7]. The state-of-art DCNN based semantic image segmentation models relies on transfering the pre-trained convolutional filters as well.[13] A typical method to pretrain the convolutional filters is to train a classification model with the large-scale ILSVRC dataset [20] However, this method constrains the semantic image segmentation models to have the same CNN architecture as the image classification models. The CNN design for semantic image segmentation does not necessarily follow the de-

[1] M: Do we actually have a reference for that? Or other proof?
J: I think the preparation methods described in these papers can be supportive. Microsoft COCO even mentioned the total working hours in the paper.

sign of image classification architectures. The segmentation models need both global and local information to predict the category and give a fine segmentation, whereas the classification models care less about local information for object localization. For instance, the presence of the max-pooling layers enable the following convolutional filters to have larger receptive fields but, at the same time, reduce the resolution of the features. Additional upsampling layers, can recover the shape of the output segmentation but cannot fully recover the information thrown away. This first-pooling-and-then-upsampling pipeline can result in coarse segmentation output [2] with non-shape boundaries and blob-like shapes. [2] Alternatively, one can also extract convolutional features with semantic segmentation tasks. But it is more difficult to collect well-annotated dataset for semantic image segmentation than for object recognition. A large number of annotated images are of great value to train sufficiently generalised representation and avoid overfitting, given the "data-hungry" nature of DCNNs. [3] Allowing the existence noisy annotations could help signicantly increase the number of annotated images and the number of training samples could compensate the impact of annotation errors. We will further discuss this in the related works.

This paragraph should summarize the main ideas.

The different levels of the hierarchical features in DCNNs are believed to play different roles in extracting information from the images. The low-level features process the local information within small neighborhood and the high-level features ensemble information from lower-level features to extract abstract information. The high-level features were found significantly dependent on the exact categories compared to the low-level features which show extraordinary category independency.[24] Previous studies[21, 17] have shown that training with noisy labels can lead to significant higher classification errors than training with clean labels if the total number of training samples are fixed. It is nevertheless unclear how the annotation errors would influence the learned multiple level features. We made a hypothesis that annotation errors do not necessarily lead to a "bad representation" because the "generality" of low-level features may contribute to a robustness to the rrors when we transfer the learned features to a new dataset with new categories.

Table of contents

In the next section, we summarized the related works. In Section 3 we formulate the annotation errors into three categroies: misannotation, misclassification and incomple annotation. We tested our hypothesis in Section 4, studying whether the misannotation and misclassification had an impact on learning "transferable" features. In Section 5 we connected training with inexaustive annotations to Positive and Unlabeled Learning.

2 Related work

Semantic Image Segmentation with Deep Neural Nets J. Long et.al.[13] defined a skip architecture to combine semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations and transfered the learned representations from the contemporray classification networks into fully convolutional networks. L. Chen et.al.[2] removed the last few max pooling layers of the CNNs and upsampled the corresponding filters to avoid the reduced feature resolution by the pooling layers. An additional fully connected Conditional Random Field (CRF) was added to refine the coarse last layer output for better localization performance. S. Zheng et.al.[25] integrate the CRFs-based probabilistic graphical modeling with CNNs in an end-to-end framework.

Transfer Learning The first-layer features in the modern CNNs are often observed converging to either Gabor filters or color blobs, regardless the exact learning objectives and the training dataset. This pheonomen is called the *generality* of the first-layer features. By contrast, the last-layer features depend significantly on the learning objective and dataset and they are called *specific*. Yosinski et.al.[24] studied the transition from general to specific for the features in the intermediate layers by measuring how much the tranformed pre-trained features boost the fine-tuning performance on a new dataset, i.e. the "transferability" of the features. They found features from several bottom layers, not only the first layer, were "transferable"

[2] M: Also for this claim, I think we need a reference or something like that. Or other proof indeed... In addition, I wonder whether we can explain why this may be the case? J: Yes, I can refer to a paragraph in the intro of CRFasRNN and enrich the discussion a bit.

[3] M: For the rest a bit vague... what do we really mean by suffer? Maybe this becomes clear later in the intro...? J: Refrased.

to a new dataset, meaning that they were not specific for a particular category but were shared across categories. This discovery led to our hypothesis that the general features can be more robust to annotation errors, either object misannotation or instance misclassidication, than the specific features. We used the same measure as Yosinki et.al. did to quantitatively measure the "transferability" of the features and reported the robustness to instance misannotation and instance misclassification for the learned representation.

Unsupervised and semi-supervised pre-training Apart from supervised pre-training, one can also obtain the pre-trained features in an unsupervised or a semi-supervised way. The most common method is to train a generative model with either *auto-encoder* variants or *deep beilief network*. Vincent et al.[22] trained multiple levels of representation robust to the corrupted inputs with stacked denoising auto-encoders. Masci et al.[14] presented a stacked convolutional auto-encoder unsupervised pre-training for hierarchical feature extraction. Hinton et al.[10] proposed a greedy learning algorithm to train *deep belief nets* one layer at a time to train hierarchival features. Lee et al.[11] presented a *convolutional deep belief network*, to learn hierachical convolutional representations. A few studies[4, 3, 1] highlighted the advantage of unsupervised pre-training compared to the random initialization, connecting unsupervised pre-training to a norm of regularization and a method that help disentangle the sample variations. However, better random initialization strategies, for example xavier initialization[8] and its variants, have shorten the gap between unsupervised pre-training and random initialization. Using unsupervised pre-training or not now becomes a tradeoff between the time and resources invested and the perfomance gain. Unsupervised deep representation learning is in general not comparable to supervised representation learning especially when large scale dataset is available. A proper method to learn features in the presence of label noise should outperform unsupervised pre-training because noisy information is still better than no information. [4] *TODO Semi-supervised representation learning*

[4] J: This argument is a bit too strong to me.

Deep Learning with Noisy Labels A few studies[21, 17] investigated the impact of label noise on classification performance with convolutional neural networks assuming the labels were randomly transited from one to another given the probabilities fall in a trainsition matrix. They found a significant decrease of classification performance along with the increase of false label proportion when the total number of examples is fixed. They then proposed methods to handle this label noise at random (NAR)[6] situation by either introducing a linear noise layer on top of the output layer[21] or correcting the loss functions with an estimation of the noise transition matrix[17]. Xiao et al.[23] integrated a probabilitic graphic model to an end-to-end deep learning system to train predicting class labels, either correct or wrong, together with correcting the wrong labels. Reed & Lee[18] proposed an empricaly way of taking into account the *perceptual consistency* for large-scale object recognition and detection when incomplete and noisy labels exist by introducing a bootstrpping modification to the negative log-likelihood, in either a "Hard" or a "soft" favor. The "soft" bootstrapping loss is actually equivalent to a softmax regression with *minimum entropy regularization*[9] which was originally proposed for semi-supervised learning. Minimum entropy regularization encourages the model to have a high confidence in predicting labels.

Noise robustness In contrast to the aforementioned works, Rolnick et al.[19] argued that deep neural networks can learn robustly from noisy dataset as long as approriate hyperparameters choice are made. They studied instead of replacing the correct labels with noisy labels but diluting correct labels with noisy labels to support their argument. They then concluded sufficiently large training set is of more importance than lower the level of noise. This work is closely related to our work in Section 4, except that we focus on the label noise robustness with respect to the feature "transferability" instead of the classification performance.

Positive and Unlabeled Learning Most of the current studies for deep learning with noisy labels focus on classification problems.

3 Problem Formulation

Briefly formulate the problem.
Semantic Segmentation tasks can be considered as per-pixel classification problem. Each of the pixels is assigned a label of either 0, indicating a background pixel, or $k \in 1, \ldots, K$ denoting a foreground pixel corresponding to an instance from one of the K categories. The aforementioned errors can be interpreted by the pixel label flipping: *misannotation* is label flipped from 0 to k, *inexaustive annotation* flipped from k to 0, and *misclassification* flipped from i to j, where $i, j, k \in 1, \ldots, K$. The misannotated instances are assumed to be visually distinguishable against the background and have natural semantic meaning. All the label flips have one instance as the minimum fliping unit.
Instance Misannotation and Instance Misclassification One hypothesis made in this work is that the *misannotated and misclassified* instances can still provide information for training multi-scaled features that are *transferable*[24] to the other datasets and categories. Supposing a dog toy is wrongly annotated as a dog, given that the "dog" is one of the target categories while the "toy" is not, the error would introduce bias into the last layer but not necessarily into the first convolutional layers. The high-level features were found more dependent on a particular category, i.e. more *specific*, than the low-level features, whereas the low-level features were found less category-dependent, i.e. more *general*. [24] We wanted to explore whether the "generality" of low-level features contributes to the annotation error robustness when we transfer the learned features to a new dataset with new categories. That leads us to the following research question:

1. How do misannotation and misclassification of instances influence the "transferability" of the learned features respectively?

We experimented, in Section 4, how transferable the learned features are in two special cases for the two types of annotation error:

instance misannotation all instances from the non-target categories are misannotated as instance from the target category

instance misclassification the classes for all the instances are completely randomly assigned

Initial Representation	mean IU (aerospace, bicycle, bird, boat, bottle)	mean IU (bus, car, cat, chair, cow)	mean IU (dining table, dog, horse, motorbike, person)	mean IU (potted plant, sheep, sofa, train, TV)
ImageNetModel	0.42 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	0.47 ± 0.01
SingleCategory	0.42 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	0.47 ± 0.01
PixelObjectness	0.30 ± 0.02	0.35 ± 0.01	0.29 ± 0.02	0.35 ± 0.03
CompleteCategory	0.29 ± 0.01	0.36 ± 0.01	0.29 ± 0.01	0.37 ± 0.01
RandomCategory	0.29 ± 0.01	0.33 ± 0.03	0.26 ± 0.01	0.28 ± 0.01
FromScratch	0.29 ± 0.01	0.29 ± 0.03	0.27 ± 0.01	0.30 ± 0.02

Table 1: Performances of FCN with Alexnet trained to segment 5 categories from the PASCAL VOC2011 dataset with different representation initializations. *CompleteCategory* is the model pre-trained to segment the other 15 categories from the PASCAL VOC2011 dataset; The *PixelObjectness* model was pre-trained to distinguish the instance against the background; The *RandomCategory* model was pre-trained to segment instances with randomly assigned categories from 1 to 15.

[5] The *transferability* of the features can be evaluated by how much they can boost the performance of training a new dataset. [24] *TODO One sentence summarizes the results.*

Inexaustive annotating

The exhaustive annotations can introduce bias to both the decoding layer and the encoding layers because they negatively contribute to the activations in all the layers. [6] The inexhaustive annotations need to be properly handled given the prior knowledge modeling the missing pattern of the annotations. Given that we believe any annotated instance provide information, all the foreground pixels that correspond to the annotated instances become reliable and the background pixels may contain both the true background pixels and object pixels unannotated. That satisfies a Positive and Unlabeled learning setup where the training dataset contains only the positive examples and unlabeled examples that are the mixed of the positive samples and negative samples.

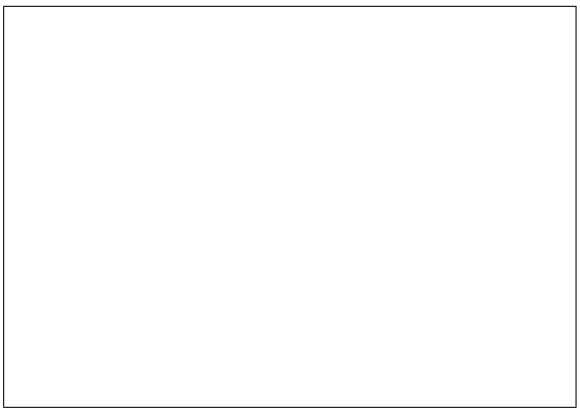

Figure 1: Visualization of first-layer features from different pre-trained models.

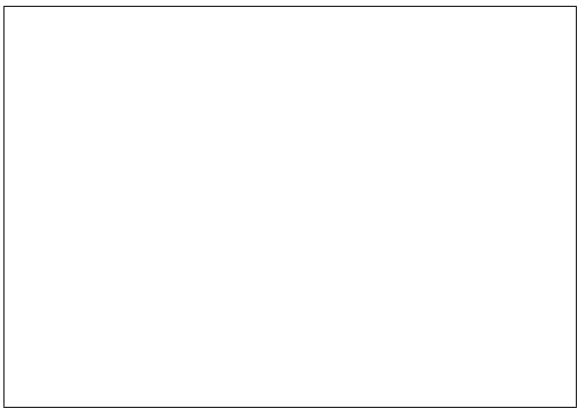

Figure 2: Varying the number of categories while pre-training the representation and the pre-trained weights were fine-tuned to segment 5 categories from the PASCAL VOC2011 dataset.

Annotation	Loss	acc.	prec.	rec.	F_1
Complete	CrossEntropyU.	0.87 ± 0.01	0.88 ± 0.01	0.82 ± 0.01	0.85 ± 0.01
50%(P+N)	CrossEntropyU.	0.83 ± 0.01	0.84 ± 0.01	0.78 ± 0.01	0.80 ± 0.01
50%P+U	CrossEntropyU.	0.64 ± 0.04	0.93 ± 0.08	0.34 ± 0.02	0.44 ± 0.06
50%P+U	WeightedU.	0.78 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.76 ± 0.01
50%P+U	ExponentialU.	0.82 ± 0.01	0.86 ± 0.01	0.73 ± 0.01	0.78 ± 0.01
50%P+U	BootstrapHard	0.74	0.81	0.60	0.67
50%P+U	DropoutReg.				

Table 2: Image classification with positive examples partially annotated. The complete dataset contains images from CIFAR10 as the **positive** (P) set and images from CIFAR110 as the **negative** (N) set. The unannotated positive examples from P set construct the **unlabeled** (U) set together with the N set.

4 Pre-train features by learning "objectness"

5 Positive and Unlabeled Learning

One sentence summary of Positive and Unlabeled Learning

Formulation *This part should explain the Positive and Unlabeled Learning setup with mathematical representation when necessary.*

Weighted Logistic Regression *This part should discuss the linear model for observing positive conditioning on true positive and its relationship to changing the class weight.*

Exponential Loss for unlabeled examples *This part should explain why the exponential loss could perform better than the cross-entropy loss, potentially with a figure of 2D Gaussians.*
This paragraph should explain why fade-in was introduced to avoid all-positive inital prediction

[5]M: So, to what extent is this the actual research question that you would like to answer? J: I think this section answers your question now.

[6]J: This argument need evidence too, or experiment/discussions in details in Section 5

This part should explain the influence of the imbalanced problem and how to overcome.

6 Results

7 Conclution

References

[1] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

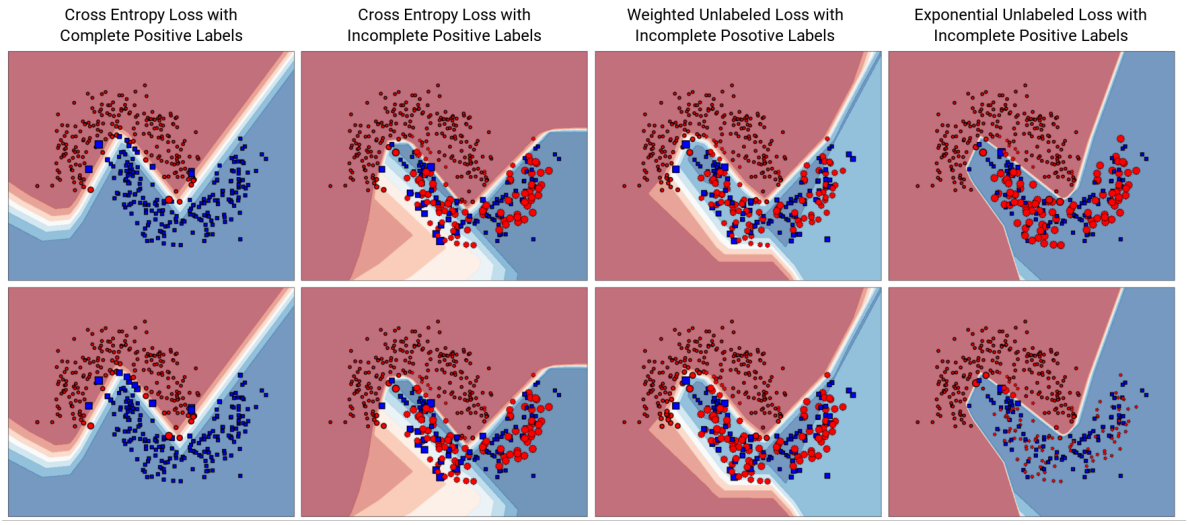

Figure 3: 2D moons dataset with non-linear separable decision boudary. Four hundreds samples per class were drawn randomly from two interleaving half circles with noises added with a minor standard deviation. A **red circle** indicates an example labelled as positive whilst a **blue square** indicates the example has a negative label. The **leftmost** figures have complete positive labels, meaning the positive and negative labels are all correct, whereas, in **the other figures** only half of the positives were correctly labelled and the rest were mixed with the negative samples. The **background colors** represent the probability for the area to be positive given by the classifier trained with the given samples and labels: **red** for high probability areas, **blue** for low probability areas and **white** for the class transition areas, i.e.decision boundaries. The **size of the markers** in the top row denotes the per-class normalized training losses and the **size of the markers** in the bottom row the per-class normalized derivatives w.r.t the output of the last layer for the trained Multilayer Perceptron (MLP) with the different losses.

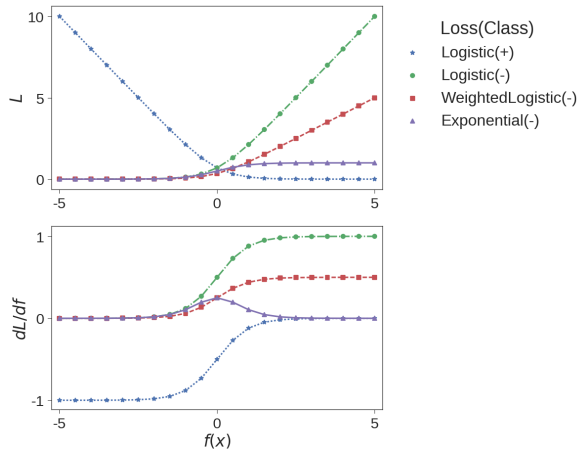

Figure 4: The Logistic Loss, Weighted Logistic Loss, Exponential Loss and their dirivatives with respect to the model output.

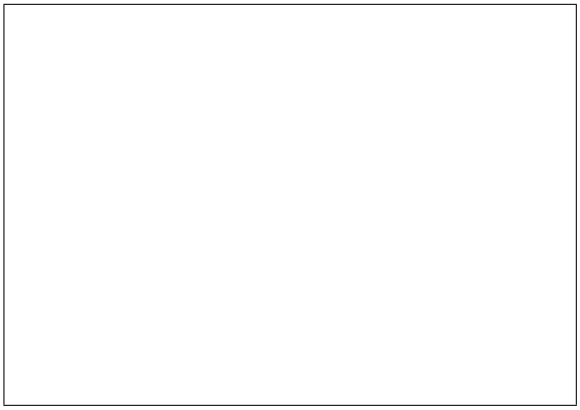

Figure 5: Varying percentage of annotated positives 10%, 20%, 50%, 80% and 100% with images from CIFAR10 as the positives and images from CIFAR110 as the negatives.

Annotation	Loss	pixel acc.	mean acc.	mean IU	f.w. IU
Complete	CrossEnt.U				
50%(P+N)	CrossEnt.U				
50%P+U	CrossEnt.U				
50%P+U	WeightedU				
50%P+U	ExponentialU				
50%P+U	BootstrapHard				
50%P+U	DropoutReg.				

Table 3: Image semantic segmentation with images contain single instance only from the PASCAL VOC2011 segmentation dataset. The complete **positive** (P) set denotes the foreground instances and the **negative** (N) set consists of the background. The unannotated instances from P set construct the **unlabeled** (U) set together with the N set.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[3] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[4] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pages 153–160, 2009.

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[6] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[9] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[10] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[11] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scal-

able unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[14] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 52–59, 2011.

[15] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 567–574, 2012.

[16] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[17] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making neural networks robust to label noise: a loss correction approach. *arXiv preprint arXiv:1609.03683*, 2016.

[18] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[19] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[21] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[22] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[23] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[24] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[25] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.