Journal of Electronics and Information Science (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2023.080602
ISSN 2371-9524 Vol. 8 Num. 6

A Deep Neural Network for Image Segmentation

Fan Yu[1,a,*], Haoran Gui[1,b], Huawei Wan[2,c]

[1]*Beijing University of Civil Engineering and Architecture (BUCEA), Beijing, China*
[2]*Satellite Application Center for Ecology and Environment, Ministry of Ecology and Environment, Beijing, China*
[a]*yufan021@126.com*, [b]*2867904409@qq.com*, [c]*wanhw@secmep.cn*
[*]*Corresponding author*

Keywords: Image Annotation, Machine Learning, Re-annotating

Abstract: Many tasks demand high-quality remote sensing image annotation products that are difficult to achieve through existing automated methods. Obtaining high-quality pixel annotations is time-consuming and laborious. This study proposes architecture with controllable correction ability that can automatically generate image annotations and allow annotators to adaptively correct previous annotations by making simple guidance information after discovering errors. This method can be applied to any convolution-based network. A training method and metric were proposed to measure the efficiency of re-annotation. We conducted experiments on the Vaihingen dataset using different base architectures and backbones. Our study shows that our training method can effectively direct the guidance module to utilize the guidance information and improve the re-annotation efficiency up to 2.53 times. In addition, more advanced architectures may give better results.

1. Introduction

Obtaining pixel-level class semantic labels is helpful for urban planning land cover detection and training a better machine learning model for other tasks. To obtain pixel-level semantic annotation information from a large number of images, scholars have proposed many semi-automatic and automatic methods. These methods provide pre-annotation based on prior knowledge to reduce the workload of pure manual annotation. As these models inevitably produce pixel-level mislabeling to varying degrees, it is necessary to review and correct the mislabeled pixels after pre-annotation.

With the development of aerospace technology, an increasing number of aerial images have been captured, and the ground object information described in the image has become clearer. However, the increase in information creates an increasing number of pixels that need to be annotated, and the more complex local texture also leads to a sharp increase in the difficulty of recognition. The method of purely manual annotation by pixels is becoming increasingly unrealistic, and people are turning to semantic segmentation models for pre-annotation.

Nevertheless, even the state-of-the-art semantic segmentation model cannot match annotators in all tasks. To obtain pixel-level class semantic labels with high accuracy, the review process after pre-annotation and the process of re-annotating misclassified pixels are crucial.

It is inefficient to use tools such as drawing to re-annotate the misclassified areas pixel-by-pixel.

To increase the efficiency of re-annotation, some methods based on spatial geometry and texture information [4], [6], [7], [14] allow for correction of mislabeling by adjusting their own parameters to adapt to perturbations such as contrast and scale changes [10].

Although the method based on spatial geometry and texture information can easily correct its own parameters to adapt to pixel-level semantic segmentation tasks in different scenes, searching for the appropriate parameters can be a time-consuming and tedious task. Aerial images have the characteristics of intra-class diversity and inter-class similarity; therefore, it is difficult to find a set of appropriate parameters in large images to distinguish each category well. Many model parameters are usually determined by practical experience and modifying some parameters, such as thresholds, can easily lead to changing the original correct annotation to the wrong annotation. Human beings can establish an intuitive impression of space, so they can easily judge where the error occurs, but it is difficult to judge how to modify parameters to correct this area. Compared with the pixel-by-pixel annotation method, the method based on spatial geometry and texture information is subject to parameters, yet human annotators are unable to translate their spatial intuition to machine parameters.

Compared with traditional machine learning algorithms such as decision trees [15], Naive Bayes [16], Support Vector Machines (SVM) [17, 18], and deep learning methods, especially the convolutional neural network (CNN) method, show excellent feature recognition ability and generalization ability in semantic segmentation tasks. Popular semantic segmentation models, such as SegNet [12] and U-Net [13], have been used in aerial image classification in many studies [21-25], and achieved good results. Deep convolutional neural networks (DCNNs) achieve better performance through multi-layer abstraction and tight integration, but at the cost of their internal parameters not being interpreted as intuitive things [1]. In contrast to the method based on spatial geometry and texture information, which can fine-tune the parameters to quickly correct the annotation, the deep convolution network must spend a long time to retrain after adjusting the hyper-parameters, which is very expensive. Furthermore, almost all deep learning models for semantic segmentation do not provide the ability to correct the error annotation.

We expect the deep learning model to be able to perform not only general semantic segmentation tasks, but also to receive and refer to the manual guidance information for the re-annotation task.

To solve the problems as mentioned above, this study proposes a new architecture called the deep guidance network (DGN). The DGN is mainly composed of a guidance module that performs before each convolution layer. When an image is input, the DGN is performed as a deep convolution network of semantic segmentation that can output annotation. When the guidance information is input as extra data, all guidance modules are activated, and the annotation of the output is based on the reference to the guidance signal. In the semantic annotation task, first, the DGN is used to predict the semantic labels of the input images. Then, a manual review is performed to find the mislabeled area and complete some simple marks in these areas, which are considered as guidance information. Finally, the input image and guidance information are fed into the DGN, which will be re-predicted with reference to the manual mark. The contributions of this study are summarized as follows.

We propose a guidance module for modeling the information updating process of internal features and guidance information. It determines whether it is activated according to whether guidance information is received. When activated, it performs pixel-level correction on the feature map output by the convolution layer reference to the guidance mask. It is placed before each convolution layer to support information updates at different scales.

Based on the guidance module, we propose a deep guidance network (DGN) that has an optional input to receive the guidance signal. When it does not receive the guidance information, the DGN outputs semantic labels that depend only on the images. When a guidance signal is received, all

guidance modules are activated to optimize the prediction labels.

A new training method that can train the DGN more effectively is proposed. This method prompts the guidance module to learn how to integrate guidance information into a feature map. The trained model can adaptively correct mislabeled annotations according to the guidance information.

The remainder of this study is organized as follows. Section II reviews related work. Section III describes the details of the proposed method. Section IV presents the experimental evaluation of this method. Section V presents the conclusions.

2. Related Works

This section reviews related work on semantic annotation. The existing methods can be divided into three categories: methods using annotation tools, traditional computer vision methods, and methods based on deep learning.

2.1 Tools for Image Annotation

Some tools attempt to reduce the number of interactions required for image annotation. ByLabel [2] detects the edge of an image and divides the edge points into multiple fragments. Users need only to select a small number of fragments recommended to form closed boundaries and quickly annotate the objects. Firefly [3] built a web-based annotation system that allows multiple users to use graphic objects online, such as lines, boxes, points, broken lines, polygons, curves, and free-form annotations, and provides clear visualization. It also proposed an interactive image segmentation method based on an elastic spline to assist in annotating and generating ground true images.

2.2 Method Based on Spatial Geometry and Texture Information

Currently, there are many traditional machine vision methods for extracting semantic information from images. Aiming at image interference problems such as local target occlusion, radiation difference, and image blur, Dai et al. [4] proposed combining multiple image descriptors to improve the accuracy and performance of template matching in an interference environment. The ground objects in remote sensing images usually contain significant geometric information and texture. Qin et al. [5] used the geometry of buildings, used a line segment detector to detect and extract the line segments in ultra-high-resolution optical images and connected them into closed polygons, and used the bidirectional shortest path algorithm to extract the contour of a single building. Super pixel algorithms such as Simple Linear Iterative Clustering (SLIC) are often used to segment images, but they do not consider the semantic information of the image context. Amiri [6] proposed a remote sensing image semantic annotation method based on a region adjacency graph (RAG). It uses SLIC to segment the image and construct a RAG graph, examining the context, spatial, and spectral information of the image region.

2.3 Deep Learning-Based Methods

Deep-learning-based methods are widely used in remote sensing image annotation. Convolution networks cannot capture long-distance spatial relationships. Zhao et al. proposed effective pyramid architecture PSPNet [26] to distinguish between confusion categories and inconspicuous classes. Ronneberger et al. proposed intuitive and effective encoder-decoder semantic segmentation architecture U-Net [13]. Zhou et al. proposed U-Net++ [27] based on a series of nested dense

convolutional blocks to bridge the semantic gap between the feature maps of the encoder and decoder. DeepLabv3 [28] designed atrous convolution modules to capture multi-scale contexts instead of using DenseCRF [29] post-processing for better performance. DeepLabv3+ [11] added a decoder module by extending DeepLabv3 to extract features at an arbitrary resolution. Mou et al. [7] proposed a spatial relationship module and a channel relationship module to learn and infer the global relationship between any two spatial locations or feature maps and generate a relationship enhancement (RA) feature representation. Niu et al. proposed the HMANet [8] architecture, which can more effectively capture global associations from the perspective of space, channel, and category. It uses the class augmented attention module embedded in the class channel attention module to calculate the class-based association, recalibrate the class-level information, and introduce a region shuffle attention module to reduce feature redundancy and improve the efficiency of the self-attention mechanism through region representation. When using a convolution network to annotate a new city that is not included in the training set, its accuracy may be significantly reduced. Benjdira et al. [9] proposed an unsupervised domain adaptive semantic segmentation method based on the generation adversities network [30-32], which can effectively improve the segmentation accuracy when migrating to a target domain captured by a different sensor is shown Figure 1.

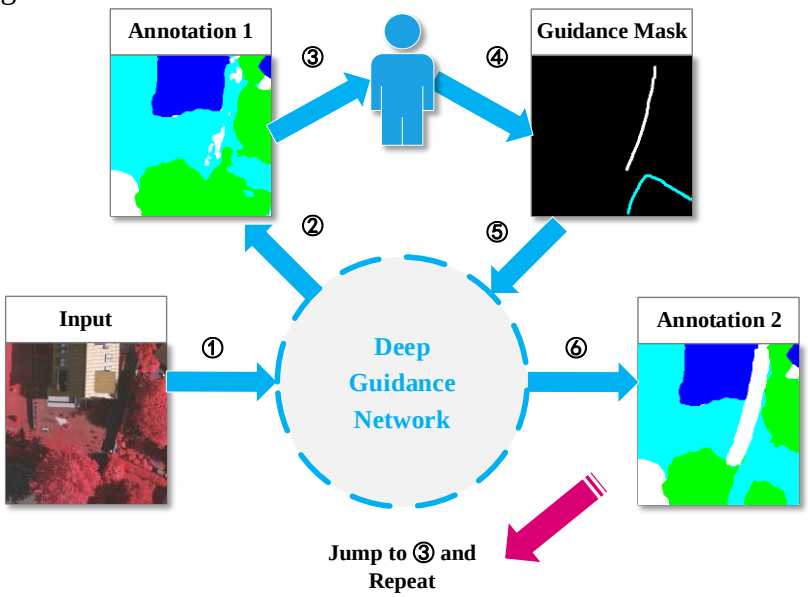

Figure 1: Pipeline of the proposed deep guidance network.

3. Methodology

This section describes the pipeline of the proposed method. There are two modes of the DGN. The first mode is to output annotations similar to other methods based on deep learning, and the second mode allows users to correct previous annotations.

Fig. 1 shows this approach. Specifically, when given an input image, it is first fed into the DGN in step (1). The DGN adaptively extracts features and provides the first annotation image in step (2). Then, the annotated image is submitted to the user for manual review in step (3). When necessary, the user can make a guidance mask in step (4) containing guidance information and feed the guidance mask and the input image into the DGN in step (5). The DGN activates the guidance module as much as possible to embed the guidance information into the internal features, and give the second annotation image in step (6), which fully refers to the guidance information. To further improve the performance of the annotation, we back to step (7).

The DGN architecture can be seen as a fusion architecture that places a guidance module before each convolutional layer within a backbone semantic segmentation architecture. In addition, each guidance module in the network is associated with a pre-processing module, which is used to pre-process the guidance mask for supporting all guidance modules in different resolution scales.

Guidance modules have no strict requirements for the backbone, but better results can be achieved with more advanced architectures. In some architectures, feature maps may be treated in a special way, such as cropping or scaling to a size that is not supported by the guidance module. In this case, the guidance modules are not all active, even if extra guidance information are input is shown Figure 2.

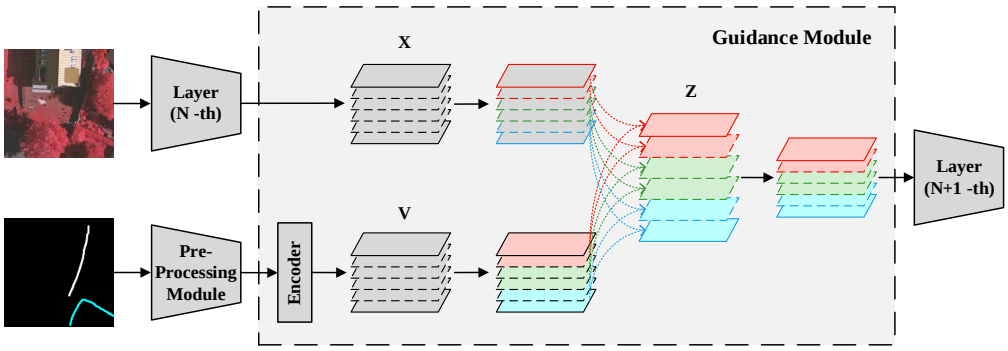

Figure 2: Pipeline of the proposed guidance module.

4. Experiments

Vaihingen is a relatively small village with many detached buildings and small multistory buildings. The Vaihingen dataset [19] contains 33 high-resolution true orthophoto (TOP) images with three spectral bands (red, green, and near-infrared) and the corresponding digital surface models (DSMs). The dataset contains six categories of semantic targets: impervious surfaces (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and background (red).

We used confusion matrices and evaluation metrics, including overall accuracy (OA), kappa coefficient (Kappa), intersection over union (IoU), and mean intersection over union (mIoU) to quantitatively assess model performance.

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{2}$$

$$Kappa = \frac{N\sum_1^t N_{kk} - \sum_1^t (N_{k\sum} X_{\sum k})}{N^2 - \sum_1^t X_{k\sum} X_{\sum k}} \tag{3}$$

TP, TN, FP, and FN are the samples of true positives, true negatives, false positives, and false negatives in the confusion matrix, k denotes a category, and t is the total number of categories; N denotes the total number of pixels; N_{kk} refers to the number of pixels correctly classified on the diagonal; $N_{k\sum}$ and $N_{\sum k}$ are the sum of the pixels in the k-th row and k-th column, respectively.

The confusion matrix changes with the input of guidance information. Positive correction (PC) is defined as the number of pixel samples converted from FP to TP (or from FN to TN). Correspondingly, negative correction (NC) is defined as the number of pixel samples converted from TP to FP (or from TN to FN). The input guidance map is based on the final output of the network. The total guidance pixel (TGP) is defined as the number of non-zero pixels in the guidance mask. The correction score (CS) was proposed to fairly compare the performance in different architecture settings with different guidance maps:

$$CS = \frac{PC-NC}{TGP} \quad (4)$$

On average, for every pixel specified in the guidance mask, the network corrects the CS pixels in the output. CS values provide a measure of the workload of handmade annotations to some extent. A higher CS value indicates that the network can use less guidance information and obtain better guidance performance.

We used pretrained models on ImageNet [32] as the backbone. The Adam optimizer [20] was used in the training, and its initial learning rate was 0.001. Beta1 and beta2 were 0.9 and 0.999, respectively. In the GM layer, λ was set to 0.95. The input image was cropped randomly from the original image. In terms of data enhancement, random horizontal flip, vertical flip, and random scaling between 0.8 ~ 1.5 magnification were applied. Finally, the input image was scaled to 384×384, the batch size was set to 8, and 20k iterations were performed.

To study the performance of our proposed method, we conducted experiments on the Vaihingen dataset using different DGN architectures and different backbones.

We begin by using some popular semantic segmentation architectures as the DGN base architecture. The GM is placed before each convolution layer and the number of accepted channels is set to be the same as the number of channels of the feature map accepted by the following convolution layer. When the number of channels is less than 16, the channel grouping k is set to 1; otherwise, k is set to the number of channels divided by 16 (unless it is not divisible by an integer).

Table 1: Results of different base architectures with same backbone ResNet-50

Method	Backbone	OA(%)	mIoU(%)	Kappa(%)	TGP(%)	PC(%)	NC(%)	CS
DNG-FCN	ResNet-50	82.21	56.04	76.63	1.09	1.76	0.56	1.11
DNG-PSPNet	ResNet-50	74.1	49.40	66.59	1.28	1.98	0.34	1.28
DNG-LinkNet	ResNet-50	79.36	53.99	70.99	0.67	1.42	0.21	1.80
DNG-U-Net	ResNet-50	80.54	57.99	75.64	1.07	2.57	0.29	2.12
DNG-U-Net++	ResNet-50	78.07	53.26	72.08	1.05	2.49	0.41	1.99
DNG-DeepLabv3	ResNet-50	78.86	54.73	73.32	0.75	1.70	0.35	1.80
DNG-DeepLabv3++	ResNet-50	82.39	61.55	78.31	0.64	1.58	0.22	2.13

The TGP/PC/NC ratio was divided by the total number of pixel samples to make it more intuitive in Table 1, and we observed the best OA and CS results from DGN-DeepLabv3++ of 82.39% and 2.13, respectively. We attribute our results to the design that DeepLabv3++'s dilated convolution layers and multi-scale fusion strategies provide ample space for guidance modules to inject guidance information. In addition, it can be noted that U-Net achieved the second highest CS value of 2.12. We believe this is because U-Net supports up to six resolutions of the guidance mask and holds more high-resolution feature maps than U-Net.

5. Conclusions

Aerial image annotation is of great value. Although existing deep learning methods can replace manual annotation, these methods do not support modifying the previous annotation when quality inspectors find the annotation error. In this study, a GM for injecting guidance information into the feature map of a deep network is proposed, and the deep guidance network is proposed, which simply places the GM in front of all convolutional layers. DGN can automatically generate aerial image annotations and allow annotators to adaptively correct previous annotations by providing simple guidance information after discovering errors. To solve the problem of how to apply

guidance information to layers with different resolutions, a downsampling method of guidance mask is proposed to support GM layer input with different scales, and a confidence based method is used to generate guidance information to effectively train the network. Finally, the network effect is verified on the Vaihingen dataset, which proves that the network can effectively combine the guidance information to correct the previous mislabeling.

Acknowledgment

This work is jointly supported by National Key Research and Development Program of China (2021YFE0194700, 2021YFB2600101) and R&D Program of Beijing Municipal Education Commission (KM202010016010, Research on the quality detection method of ground cover classification based on deep confidence neural network).

References

[1] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
[2] Qin, Xuebin, et al. "Bylabel: A boundary based semi-automatic image annotation tool." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.
[3] Sampathkmar, Urmila, et al. "Assisted ground truth generation using interactive segmentation on a visualization and annotation tool." 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2016.
[4] Dai, Jiguang, et al. "Road extraction from high-resolution satellite images based on multiple descriptors." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020): 227-240.
[5] Qin, X., et al. "Accurate Outline Extraction of Individual Building from Very High-Resolution Optical Images." IEEE Geoscience and Remote Sensing Letters 15.11(2018):1775-1779.
[6] Amiri, Khitem, and Mohamed Farah. "Graph of concepts for semantic annotation of remotely sensed images based on direct neighbors in RAG." Canadian Journal of Remote Sensing 44.6 (2018): 551-574.
[7] Mou, Lichao, Yuansheng Hua, and Xiao Xiang Zhu. "Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High-Resolution Aerial Images." IEEE Transactions on Geoscience and Remote Sensing 58.11 (2020): 7557-7569.
[8] Niu, Ruigang, et al. "Hybrid multiple attention network for semantic segmentation in aerial images." IEEE Transactions on Geoscience and Remote Sensing (2021).
[9] Benjdira, Bilel, et al. "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images." Remote Sensing 11.11 (2019): 1369.
[10] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.
[11] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.
[12] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.
[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.
[14] Xu, Lu, et al. "Farmland extraction from high spatial resolution remote sensing images based on stratified scale pre-estimation." Remote Sensing 11.2 (2019): 108.
[15] Phiri, Darius, et al. "Decision tree algorithms for developing rulesets for object-based land cover classification." ISPRS International Journal of Geo-Information 9.5 (2020): 329.
[16] Memon, Nimrabanu, Samir B. Patel, and Dhruvesh P. Patel. "A Novel Approach of Polsar Image Classification Using Na ̈ve Bayes Classifier." Mathematical Modeling, Computational Intelligence Techniques and Renewable Energy: Proceedings of the First International Conference, MMCITRE 2020. Springer Singapore, 2021.
[17] Rana, Vikas Kumar, and Tallavajhala Maruthi Venkata Suryanarayana. "Performance evaluation of MLE, RF and SVM classification algorithms for watershed scale land use/land cover mapping using sentinel 2 bands." Remote Sensing Applications: Society and Environment 19 (2020): 100351.
[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Pro-ceedings of the IEEE conference on computer vision and pat-tern recognition, pages 3431–3440, 2015.

[19] Isprs.2d semantic labeling contest-vaihingen. [Online].Available: http://www2.isprs.org/ commissions/ comm3/ wg4/ 2d-sem-label-vaihingen.html
[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint, 2014.
[21] He, Nanjun, Leyuan Fang, and Antonio Plaza. "Hybrid first and second order attention Unet for building segmentation in remote sensing images." Science China Information Sciences 63.4 (2020): 1-12.
[22] Abdollahi, Abolfazl, Biswajeet Pradhan, and Abdullah M. Alamri. "An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images." Geocarto International (2020): 1-16.
[23] Lin, Yeneng, et al. "Road extraction from very-high-resolution remote sensing images via a nested SE-Deeplab model." Remote sensing 12.18 (2020): 2985.
[24] Hou, Yuewu, et al. "C-UNet: Complement UNet for Remote Sensing Road Extraction." Sensors 21.6 (2021): 2153.
[25] Campos, Adrian, et al. "Deep Convolutional Neural Networks for Road Extraction." 2020 IEEE Green Energy and Smart Systems Conference (IGESSC). IEEE, 2020.
[26] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[27] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018. 3-11.
[28] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv: 1706. 05587 (2017).
[29] Krähenbühl, Philipp, and Vladlen Koltun. "Efficient inference in fully connected crfs with gaussian edge potentials." Advances in neural information processing systems 24 (2011): 109-117.
[30] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in CVPR, 2017.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bern-stein et al., "Imagenet large scale visual recognition challenge, "International journal of computer vision, 115(3), 211–252 (2015).