Unsupervised Deep Feature Transfer for Low Resolution Image Classification

Yuanwei Wu[1]*, Ziming Zhang[2]†, and Guanghui Wang[1]
[1] EECS, The University of Kansas, Lawrence, KS 66045
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139
y262w558@ku.edu, zzhang@merl.com, ghwang@ku.edu

Abstract

In this paper, we propose a simple while effective unsupervised deep feature transfer algorithm for low resolution image classification. No fine-tuning on convenet filters is required in our method. We use pre-trained convenet to extract features for both high- and low-resolution images, and then feed them into a two-layer feature transfer network for knowledge transfer. A SVM classifier is learned directly using these transferred low resolution features. Our network can be embedded into the state-of-the-art deep neural networks as a plug-in feature enhancement module. It preserves data structures in feature space for high resolution images, and transfers the distinguishing features from a well-structured source domain (high resolution features space) to a not well-organized target domain (low resolution features space). Extensive experiments on VOC2007 test set show that the proposed method achieves significant improvements over the baseline of using feature extraction.

1. Introduction

Recently, deep neural networks (DNNs) have demonstrated impressive results in image classification [19, 14, 3], object detection [9, 26, 22, 33], instance segmentation [12], depth estimation [15, 16], and face recognition [4]. The success of DNNs has become possible mostly due to a large amount of annotated datasets [7], as well as advances in computing resources and better learning algorithms [10, 32]. Most of these works typically assume that the images are of sufficiently high resolution (*e.g.* 224×224 or larger).

The limitation of requiring large amount of data to train DNNs has been alleviated by the introduction of transfer learning techniques. A common way to make use of transfer learning in the context of DNNs is to start from a pre-trained model in a similar task or domain, and then finetune the parameters to the new task. For example, the pre-trained model on ImageNet for classification can be finetuned for object detection on Pascal VOC [9, 26].

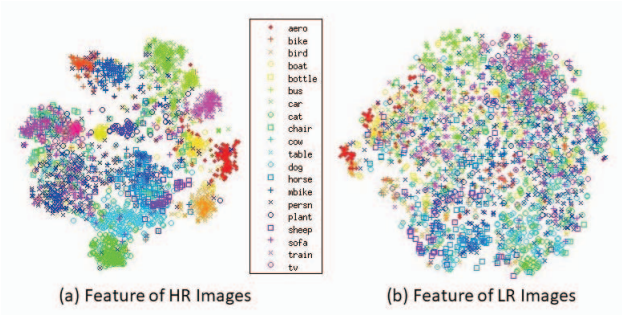

Figure 1. The tSNE [23] of deep features (2048-D) of VOC2007 train set extracted from pool5 layer of pre-trained resnet-101 [14]. (a) Feature of High Resolution (HR) images, and (b) feature of Low Resolution images. The HR features are well separated, however, the LR features are mixed together.

In this paper, we focus on low resolution (*e.g.* 32×32 or less) image classification as for privacy purpose, it is common to use low resolution images in real-world applications, such as face recognition in surveillance videos [34]. Without additional information, learning from low resolution images always reduces to an ill-posed optimization problem, and achieves a much degraded performance [25].

As shown in Fig. 1, the deep feature of high resolution images extracted from pre-trained convenet has already learned discriminative per-class feature representation. Therefore, it is able to be well separated in the tSNE visualization. However, the extracted feature of low resolution images is mixed together. A possible solution is to exploit the transfer learning, leveraging the discriminative feature representation from high resolution images to low resolution images.

In this paper, we propose a simple while effective unsupervised deep feature transfer approach that boosts classification performance in low resolution images. We assume that we have access to high resolution labeled images during training, but at test we only have low resolution images. Most existing datasets are high resolution. Moreover, it is much easier to label subcategories in high resolution images. Therefore, we

*This work was done when the first author took internship at MERL.
†Corresponding author.

2473-9944/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCVW.2019.00136

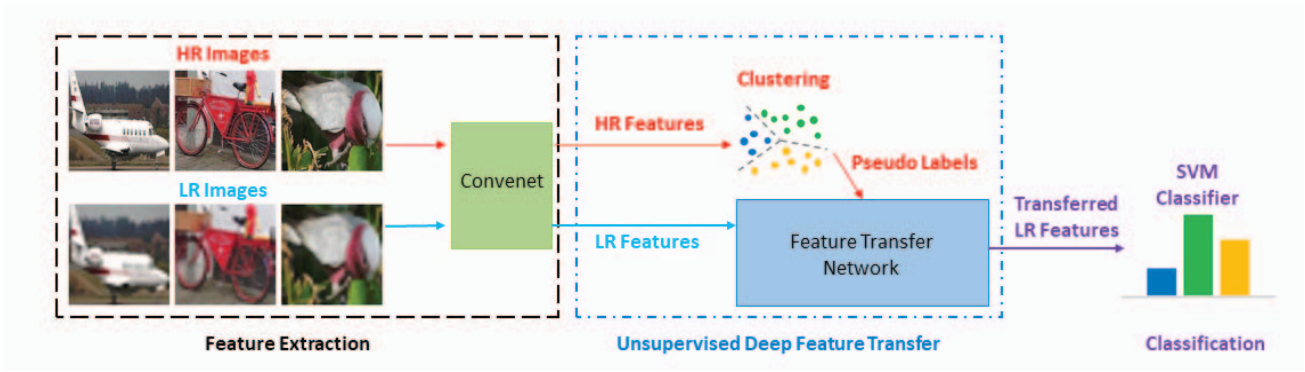

Figure 2. The overview of proposed unsupervised deep feature transfer algorithm. It consists of three modules. In the feature extraction module, a pre-trained deep convenet is used as feature extractor to obtain HR and LR features from HR and LR images, respectively. Then, we cluster the HR features to obtain pseudo-labels, which are used to guide the feature transfer learning of LR features in the feature transfer network. Finally, a SVM classifier is trained on the transferred LR features.

believe it is a reasonable assumption. We aim to transfer knowledge from such high resolution images to real world scenarios that only have low resolution images. The basic intuition behind our approach is to utilize high quality discriminative representations in the training domain to guide feature learning for the target low resolution domain.

The contributions of our work have three-fold.

- No fine-tuning on convenet filters is required in our method. We use pre-trained convenet to extract features for both high resolution and low resolution images, and then feed them into a two-layer feature transfer network for knowledge transfer. A SVM classifier is learned directly using these transferred low resolution features. Our network can be embedded into the state-of-the-art DNNs as a plug-in feature enhancement module.
- It preserves data structures in feature space for high resolution images, by transferring the discriminative features from a well-structured source domain (high resolution features space) to a not well-organized target domain (low resolution features space).
- Our performance is better than that of baseline using feature extraction approach for low resolution image classification task.

2. Related Work

Our method is closely related to unsupervised learning of features and transfer learning.

Unsupervised learning of features: Clustering has been widely used for image classification [2, 30, 17]. Ji *et al*. [17] propose invariant information clustering relying on statistical learning by optimising mutual information between related pairs for unsupervised image classification and segmentation. Caron *et al*. [2] present a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. Yang *et al*. [30] propose an approach to jointly learn deep representations and image clusters by combining agglomerative clustering with CNNs and formulate them as a recurrent process.

Transfer learning: It is commonly used in the scenario where the training and testing data distributions are different. Saenko *et al*. [28] learn a regularized non-linear transformation in the context of object recognition to minimize the effect of domain-induced changes in the feature distribution. Chen *et al*. [6] transfer knowledge stored in one previous network into each new deeper or wider network to accelerate the training of a significantly larger neural network. Yosinski *et al*. [31] experimentally study the transferability of hierarchical features in deep neural networks. Azizpour *et al*. [1] investigate the factors of transferability of a generic deep convolutional networks such as the network architecture, distribution of the training data, etc. Tzeng *et al*. [29] learn a CNN architecture to optimize domain invariance and transfer information between tasks. Long *et al*. [21] propose a deep adaptation network architecture to match the mean embeddings of different domain distributions in a reproducing kernel Hilbert space. Guo *et al*. [11] propose an adaptive fine-tuning approach to find the optimal fine-tuning strategy per instance for the target data. Readers can refer to [24] and the references therein for details about transfer learning.

3. Proposed Approach

This section describes the proposed unsupervised deep feature transfer approach.

3.1. Preliminary

With the recent success of deep learning in computer vision, the deep convnets have become a popular choice for representation learning, to map raw images to an embedding vector space of fixed dimensionality. In the context of supervised learning, they could achieve better performance than

humanbeings on standard classification benchmarks [13, 19] when trained with large amount of labelled data.

Let f_θ denote the convenet mapping function, where θ is the corresponding learnable parameters. We refer to the vector obtained by applying this mapping to an image as feature or features. Given a training set $X = \{x_1, \cdots, x_N\}$ of N images, and the corresponding ground truth labels $Y = \{y_1, \cdots, y_N\}$, we want to find an optimal parameter θ^* such that the mapping f^*_θ predicts good general features. Each image x_i associates with a class label y_i in $\{0, 1\}^k$. Let g_w denote a classifier with parameter ω. The classifier would predict the labels on top of the features $f_\theta(x_i)$. The parameter θ of the mapping function and the parameter ω of the classifier are then learned jointly by optimizing the following objective function:

$$\min_{\theta,\omega} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(g_w(f_\theta(x_i), y_i)), \tag{1}$$

where $\mathcal{L}$ is the multinominal logistic loss for measuring the difference between the predicted labels and ground-truth labels given training data samples.

3.2. Unsupervised Deep Feature Transfer

The idea of this work is to boost the feature learning for low resolution images by exploiting the capability of unsupervised deep feature transfer from the discriminative high resolution feature. The overview of proposed approach is shown in Fig. 2. It consists of three modules: feature extraction, unsupervised deep feature transfer, and classification, discussed below.

Feature extraction. We observe that the deep features extracted from convenet could generate well separated clusters as shown in Fig. 1. Therefore, we introduce the transfer learning to boost the low resolution features learning via the supervision from high resolution features. Then, we extract the features (N-Dimensional) of both high and low resolution images from a pre-trained deep convenet. More details are described in Sec. 4.2.

Unsupervised deep feature transfer. We propose a feature transfer network to boost the low resolution features learning. However, in our assumption, the ground truth labels for low resolution images are absent. Therefore, we need to make use of the information from high resolution features. In order to do this, we propose to cluster the high resolution features and use the subsequent cluster assignments as "pseudo-label" to guide the learning of feature transfer network with low resolution features as input. Without loss of generality, we use a standard clustering algorithm, k-means. The k-means takes a high resolution feature as input, in our case the feature $f_\theta(x_i)$ extracted from the convenet, and clusters them into k distinct groups based on a geometric criterion. Then, the pseudo-label of low resolution features are assigned by finding its nearest neighbor to the k centroids of high resolution features. Finally, the parameter of the feature transfer network is updated by optimizing Eq. (1) with mini-batch stochastic gradient descent.

Classification. The final step is to train a commonly used classifier such as Support Vector Machine (SVM) using the transferred low resolution features. In testing, given only the low resolution images, first, our algorithm extracts the features. Then feeds them to the learned feature transfer network to obtain the transferred low resolution features. Finally, we run SVM to get the classification results directly.

4. Experiments

4.1. Dataset

We conduct the low resolution classification on the PASCAL VOC2007 dataset [8] with 20 object classes. There are $5,000$ images in VOC2007 trainval set and $4,952$ images in VOC2007 test set. However, the images in the dataset are high resolution images only. We follow [20] to generate the low resolution images. In this work, we generate high resolution images by resizing the original images to 224×224 using bicubic interpolation. We generate the low resolution images by down-sampling the original to 32×32, and then up-sampling to 224×224.

4.2. Implementation Details

We conduct our experiment using Caffe [18]. We use the resnet-101 [14] pre-trained on ILSVRC2012[1] [27] as the backbone convenet to extract the features from high and low resolution images. We extract the features from the pool5 layer, which gives a feature vector with dimension of $N = 2048$.

The feature transfer network is a two-layer fully connected network. We conduct grid search to find the optimal design for the network architecture, see Sec. 4.3. It is initialized using MSRA [18] initialization. We train the feature transfer network using stochastic gradient descent with weight decay 0.0005, momentum 0.9, batch size $1,000$, epoch $1,000$, total iteration $31,561$. The initial learning rate is 0.01, and is decreased by 10 after every $15,000$ iterations.

4.3. Feature Transfer Network

The feature transfer network is shallow, with two fully connected layers. Let N_1 and N_2 denote the neurons of the first and second fully connected layers, respectively. We conduct grid search to find the optimal combination for N_1 and N_2, as shown in Table 2. The number N_2 is determined by the number of clusters k for the pseudo labels in k-means.

As we can see, when the neurons of N_2 is fixed, the mAP increases as the neurons of N_1 increases. This is because the

[1]We download the Caffe Model from https://github.com/BVLC/caffe/wiki/Model-Zoo

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
Baseline-HR	97.6	92.7	89.2	85.8	90.6	87.5	96.2	94.3	81.4	83.3	80.0	86.9	84.2	90.0	95.4	95.0	88.3	71.6	96.0	95.9	89.1
Baseline-LR	87.5	84.8	77.5	77.4	80.4	76.5	90.6	72.1	75.1	72.9	69.5	65.0	71.7	73.9	92.8	90.8	78.3	48.6	83.3	92.3	78.1
Ours	89.1	86.5	80.1	78.1	79.6	77.4	92.4	75.4	79.4	73.2	72.5	68.5	74.0	77.1	95.0	91.9	77.6	53.4	86.1	92.5	80.0

Table 1. Per-class average precision (%) for object classification on the VOC2007 test set.

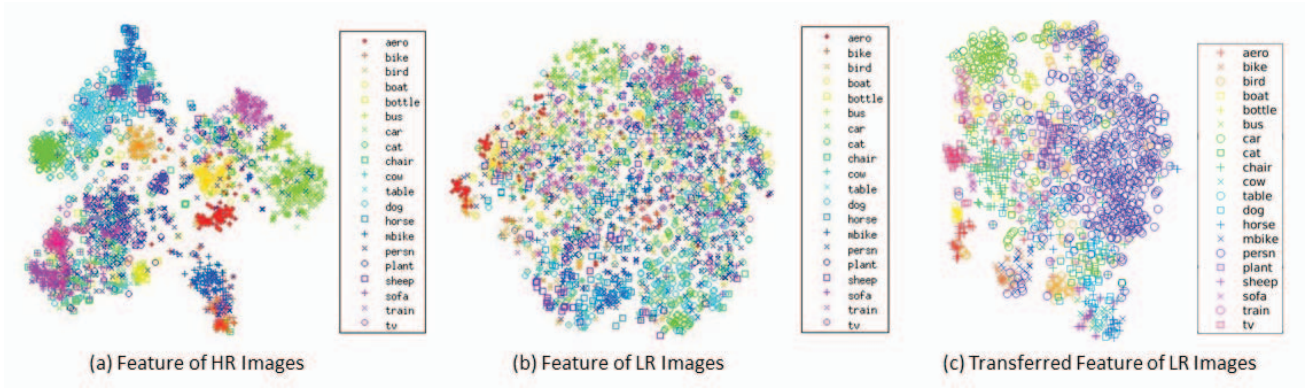

Figure 3. The tSNE of features on VOC2007 test set. (a) Feature (2048-D) of High Resolution (HR) images, (b) feature (2048-D) of Low Resolution (LR) images, (c) transferred feature (100-D) of LR image.

N_2 \ N_1	256	512	1024	2048	4096
20	0.704	0.741	**0.771**	0.786	**0.800**
100	0.718	**0.752**	0.768	**0.789**	**0.800**
200	**0.727**	0.746	**0.771**	0.788	**0.800**
500	0.717	0.743	0.766	0.784	0.795
1000	0.713	0.743	0.762	0.783	0.793
2048	0.718	0.739	0.765	0.783	0.794

Table 2. We use grid search to find the optimal combination of N_1 and N_2 for the two-layer feature transfer network by calculating the mean average precision (mAP) on VOC2007 test set.

capacity of the two-layers feature transfer network increases as the neurons increases in N_1. However, given a fixed number of neurons of N_1, the value of mAP would increase first, and then decrease when the value of neurons in N_2 is larger enough, maybe 200 is a threshold value in our two-layer network as shown in the table. We observe that the hyperparameters with $N_2 = 100$ and $N_1 = 4096$ for the neurons give us the best performance. We use the same values in our experiment.

4.4. Low Resolution Image Classification

We evaluate the performance of image classification in the context of binary classification task on the VOC2007 test set using SVM [5] classifier in matlab. We have compared our algorithm with two baselines: Baseline-HR and Baseline-LR, discussed below. Baseline-HR is to use the extracted high resolution features (2048-D) of VOC2007 trainval set to train the SVM and report the classification performance on VOC2007 test set. It is similar for Baseline-LR, but with the extracted low resolution features (2048-D). Our method transfers the low resolution feature from 2048-D to 100-D. Therefore, we train the SVM using the 100-D features for each class. We show the comparison in Table 1.

The Baseline-HR is the upper bound of our method, and Baseline-LR is the lower bound. As we can see from the Table 1, the proposed unsupervised deep feature transfer is able to boost the low resolution image classification by about 2%. Except for the classes of "bottle" and "sheep", our method outperforms the Baseline-LR. As shown in Fig. 3, we find the transferred low resolution features are sperated much better than the extracted low resolution features. Those indicate that the proposed unsupervised deep feature transfer algorithm does help transfer more discriminative representations from high resolution features. Therefore, it boost on low resolution images classification task. The feature transfer network could also be embedded into the state-of-the-art deep neural networks as an plug-in module to enhance the learned features.

5. Conclusion

In this paper, we propose an unsupervised deep feature transfer algorithm for low resolution image classification. The proposed two-layer feature transfer network is able to boost the classification by 2% on mAP. It can be embedded into the state-of-the-art deep neural networks as a plug-in feature enhancement module. While our current experiments focus on generic classification, we expect our feature enhancement module to be very useful in detection, retrieval, and category discovery settings as well in the future.

Acknowledgment

Dr. Zhang was supported by MERL. Mr. Wu and Prof. Wang were supported in part by NSF NRI and USDA NIFA under the award no. 2019-67021-28996 and KU General Research Fund (GRF).

References

[1] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation. *IEEE TPAMI*, 38(9):1790–1802, 2016. 2

[2] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 2

[3] F. Cen and G. Wang. Boosting occluded image classification via subspace decomposition-based estimation of deep features. *IEEE Transactions on Cybernetics*, pages 1–14, 2019. 1

[4] F. Cen and G. Wang. Dictionary representation of deep features for occlusion-robust face recognition. *IEEE Access*, 7:26595–26605, 2019. 1

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *TIST*, 2(3):27, 2011. 4

[6] T. Chen, I. Goodfellow, and J. Shlens. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. Ieee, 2009. 1

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, Jan. 2015. 3

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587, 2014. 1

[10] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[11] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris. Spottune: transfer learning through adaptive fine-tuning. In *IEEE CVPR*, pages 4805–4814, 2019. 2

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE ICCV*, pages 2961–2969, 2017. 1

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE ICCV*, pages 1026–1034, 2015. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 1, 3

[15] L. He, G. Wang, and Z. Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27(9):4676–4689, 2018. 1

[16] L. He, M. Yu, and G. Wang. Spindle-net: Cnns for monocular depth inference with dilation kernel method. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2504–2509. IEEE, 2018. 1

[17] X. Ji, J. F. Henriques, and A. Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. *arXiv preprint arXiv:1807.06653*, 2018. 2

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678. ACM, 2014. 3

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 1, 3

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3

[21] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 2

[22] W. Ma, Y. Wu, Z. Wang, and G. Wang. Mdcn: Multi-scale, deep inception convolutional neural networks for efficient object detection. In *ICPR*, pages 2510–2515. IEEE, 2018. 1

[23] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 1

[24] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 2

[25] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *NeurIPS*, pages 1990–1998, 2015. 1

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 1

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 3

[28] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. *ECCV*, pages 213–226, 2010. 2

[29] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE ICCV*, pages 4068–4076, 2015. 2

[30] J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *IEEE CVPR*, pages 5147–5156, 2016. 2

[31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014. 2

[32] Z. Zhang, Y. Wu, and G. Wang. Bpgrad: Towards global optimality in deep learning via branch and pruning. In *IEEE CVPR*, June 2018. 1

[33] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, Q. Nie, H. Cheng, C. Liu, X. Liu, et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[34] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE TIP*, 21(1):327–340, 2011. 1