2018 IEEE Winter Conference on Applications of Computer Vision

Ensemble Knowledge Transfer for Semantic Segmentation

Ishan Nigam Chen Huang Deva Ramanan

{inigam, chenh2, deva}@cs.cmu.edu

Carnegie Mellon University

Abstract

Semantic segmentation networks are usually learned in a strictly supervised manner, i.e., they are trained and tested on similar data distributions. Performance drops drastically in the presence of domain shifts. In this paper, we explore methods for learning across train and test distributions that dramatically differ in scene structure, viewpoints, and objects statistics. Motivated by the proliferation of aerial drone robotics, we consider the target task of semantic segmentation from aerial viewpoints. Inspired by the impact of Cityscapes [11], we introduce AeroScapes, a new dataset of 3269 images of aerial scenes (captured with a fleet of drones) annotated with dense semantic segmentations. Our dataset differs from existing segmentation datasets (that focus on ground-view or indoor-scene domains) in terms of viewpoint, scene composition, and object scales. We propose a simple but effective approach for transferring knowledge from such diverse domains (for which considerable annotated training data exists) to our target task. To do so, we train multiple models for aerial segmentation via progressive fine-tuning through each source domain. We then treat these collections of models as an ensemble that can be aggregated to significantly improve performance. We demonstrate large absolute improvements (8.12%) over widely-used standard baselines.

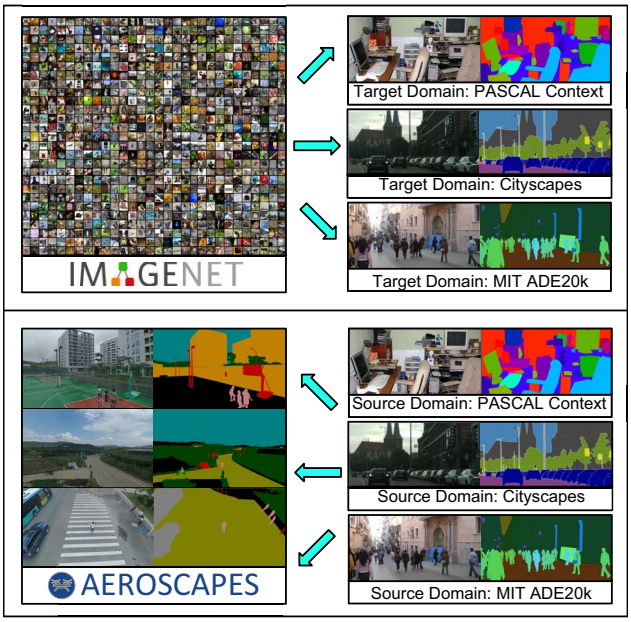


Figure 1. Contemporary recognition systems make use of multi-*target* knowledge transfer (**top**), where knowledge from a single source domain is transferred to multiple target domains. We explore multi-*source* knowledge transfer (**bottom**), where knowledge from multiple source domains is transferred to a single target domain. We propose an ensemble progressively fine-tuned via diverse source domains. We explore such issues through the illustrative task of semantic segmentation on aerial drone images and introduce *AeroScapes* - the aerial counterpart to autonomous vehicle segmentation benchmarks.

1. Introduction

Pixel-level semantic segmentation of natural scenes is a fundamental visual recognition task. Recent history has shown significant progress on standard segmentation benchmarks, *e.g.*, PASCAL VOC and Microsoft COCO [13, 29]. This success is largely owed to convolutional networks [50, 28, 8]. The community has also explored segmentation tasks that incorporate both *stuff* (amorphous background regions like grass and sky) and *things* (objects like car and person) categories [11, 51]. Other applications are found in domains such as biomedical imaging [45, 10, 1], and satellite imaging [23, 22, 33]. In particular, autonomous driving has witnessed significant development [46, 40, 2] together with an increasing number of available benchmarks [11, 42, 37].

Segmentation benchmarks: Classic semantic segmentation benchmarks have focused on general scenes, including indoor and outdoor settings [13, 29, 51, 36]. Spurred by the introduction of novel sensors, many segmentation bench-

978-1-5386-4886-5/18/$31.00 ©2018 IEEE
DOI 10.1109/WACV.2018.00168

marks have focused on limited viewpoints of specialized scenes such as ground-views of urban environments (for autonomous vehicles) [36, 11, 51], and direct overhead views (for orbital satellites) [41, 34, 23]. However, recent advances in aerial robotics allow for significantly more ease in capturing diverse viewpoints and scenes. These represent a considerable departure in statistics compared to previously-studied domains, which is the focus of our work.

Domain shift: Most deep segmentation models are deliberately trained and tested on similar data domains to attain high accuracy. Drastic performance drop is often observed in the presence of domain shifts. Indeed, domain shifts across dataset distributions pose a major challenge for learning good representations that can generalize well to all domains. Interestingly, another perspective is that *multi-source learning* of representations from such diverse source domains may, in fact, *help* generalization because each domain provides complementary information for the target task. In our work, we introduce a simple approach for transferring appropriate information from a diverse set of source domains for a particular target task.

Knowledge transfer: We turn to transfer learning techniques that allow us to transfer knowledge from existing domains (for which ample annotated data exists) to the aerial setting (for which limited annotated data exists). While transfer learning from a source to target task is a well studied problem [38, 49], by far the most common approach is fine-tuning a model pre-trained on the source task [18]. Indeed, virtually *every* contemporary visual recognition system transfers knowledge from ImageNet [43] to the target task of interest. We use this methodology to produce a ully-convolutional network (FCN) as an initial baseline, by fine-tuning on a modest set of aerial training images (e.g., ImageNet $\rightarrow$ AeroScapes). However, we would like to transfer knowledge from *multiple* domains, including indoor scenes and ground-view images of urban environments (see Fig. 1). Such source domains with richly annotated datasets represent a rich knowledge source that we would like to exploit. But the precise *manner* in which this knowledge should be transferred can be distinct and subtle - some indoor objects (such as people) can appear outdoors, and perhaps some outdoor objects look similar under aerial viewpoints (such as bicycles and motorcyles).

Ensemble transfer: Our key insight is to combine knowledge from multiple sources by learning an *ensemble* of models that are trained with *progressive fine-tuning* (ImageNet $\rightarrow$ PASCAL $\rightarrow$ AeroScapes, ImageNet $\rightarrow$ Cityscapes $\rightarrow$ AeroScapes, etc.). Intuitively, each model in the ensemble makes use of different source knowledge and so will likely make different errors (e.g., PASCAL models may be more accurate on people because they occur often in PASCAL, while Cityscapes models may be more accurate on vehicles). We then optimally combine these ensembles so as to obtain a final prediction. Our ensemble model improves over strong baselines by 8.12%. In summary, the contributions of this research is as follows:

- We propose a novel architecture-agnostic method to transfer knowledge present in diverse data sources, as encoded by richly-labeled source datasets tailored for domains *other* than the target domain of interest.
- We release the AeroScapes aerial semantic segmentation dataset, captured to study transferability of knowledge from multiple segmentation benchmarks.
- We experimentally validate our proposed benchmark using Fully Convolutional Networks, and report significant improvements over strong baselines trained with widely-adopted best-practices.

2. Related Work

Semantic segmentation: Start-of-the-art semantic segmentation methods use the convolutional networks to learn a pixel-to-pixel mapping from the image space to semantic label space [30, 9, 50, 28, 27, 12]. The success of these deep neural networks can be attributed to the availability of a large amount of pixel-level annotations and the ability of deep nets to learn from large data in an end-to-end manner. One of the most successful deep models is the Fully Convolutional Network (FCN) [30] that can directly generate the spatial label map as output.

Multi-task learning: Multi-task learning improves model generalization by combining domain-specific information learned through complementary tasks on each domain [6]. These methods usually learn a generalizable representation by learning representations across domains. Inspired by the *multi-task learning* paradigm, we present a *multi-source learning* framework, which learns a representation for a single target domain from multiple source representations. Theoretically, it is possible to learn a single representation from different domains under a multi-task framework [25]. However, in practice, this requires appropriate weighting among different tasks and a large memory budget to deal with multi-domain data simultaneously. Our proposed multi-source learning framework proves to achieves competitive results in a simple but effective way.

Knowledge Transfer: Pixel-level annotation of semantic categories is a time consuming endeavor. A rich literature employs semi-supervised and weakly-supervised learning methods to aid such tedious labelling efforts, which can be regarded as knowledge transfer in the label space. Weak supervision is generally provided as class-level labels

Figure 2. The AeroScapes Semantic Segmentation Dataset captures aerial outdoor scenes using a drone. The dataset comprises of 3269 images and ground truth segmentation maps for both *stuff* and *thing* categories.

[24], specific point annotations [3], object localizations [48], or saliency mechanisms [21].The authors in [39] develop an Expectation-Maximization framework for image segmentation under both weakly-supervised and semi-supervised settings. Recently, Chaudhry et al. [7] combined the saliency and attention maps to obtain reliable cues to boost segmentation performance and effectively explore knowledge from class labels.

Domain Adaptation: Domain adaptation methods aim to address the gap between the distributions across different data domains [26]. Recent deep learning-based methods align the domain features by maximizing the confusion [14, 15, 47] or explicitly minimizing the distances [31, 32] between their distributions across domain. To our knowledge, [19] is the only deep domain adaptation method applied to semantic segmentation. It involves image domain adversarial training and class distribution alignment, which renders learning difficult. Many domain adaptation methods focus on scenarios where little or no labeled data is available for the target domain. In our case, we have put forth considerable effort to collect and annotate the AeroScapes dataset, and so use the well-established paradigm of fine-tuning to transfer knowledge from multiple source domains to our target AeroScapes domain.

3. AeroScapes Semantic Segmentation Dataset

Most classical localization benchmarks focus on understanding objects in images, disregarding the setting in which the objects occur. Background elements provide semantic and geometric context for objects in the foreground [36, 5]. For example, an autonomous car may navigate based on roads it identifies in its line of sight, or the path planner may require that the car never attempts to park on sky or water. Thus, it is imperative that terrain-based or aerial autonomous agents are taught to identify both foreground as well as background elements.

The ability to foresee events in the future is a critical attribute of real-time autonomous systems, which rely on scene understanding for decision making. An appropriate test bed for such systems must incorporate labeled image sequences [42, 11]. Agents that rely on visual scene understanding for decision making must also learn to incorporate temporal information into their representations. Thus, it is necessary that evaluation benchmarks for navigation systems incorporate video data.

Aerial robots allow us capture previously unexplored viewpoints and diverse environments. While autonomous cars are constrained to move on the ground, aerial robots have the freedom to navigate in three-dimensions, allowing us to capture visual scales and view-points that are richer and more varied than prior benchmarks. The above constraints motivate us to collect the AeroScapes Dataset [1], which contains images captured from a drone operating at an altitude of 5-50 meters. The segmentation maps associated with these images are labeled with both *stuff* classes - vegetation, roads, sky, construction - and *thing* classes - person, bikes, cars, drones, boats, obstacles, animals (Fig.2).

The AeroScapes dataset comprises of 3269 images acquired from 141 video sequences, and contains several video sequences that are temporally downsampled. The class distribution in AeroScapes reflects the data imbalance observed in typical outdoor images comprising of both stuff and things annotations. The cumulative weight of the things classes is approximately 1.51% of the data (Fig. 3).

[1] AeroScapes Dataset: http://www.github.com/ishann/aeroscapes

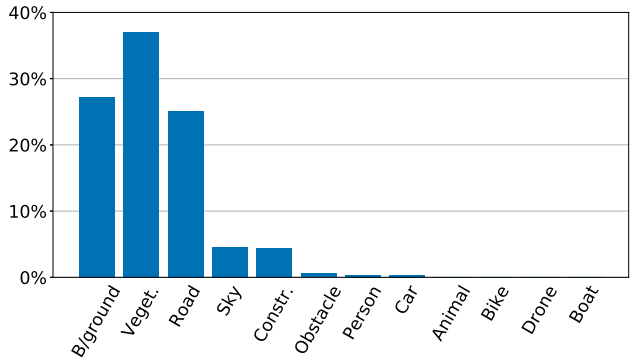
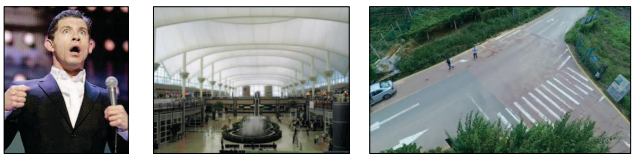


Figure 3. Pixel distribution of the AeroScapes Dataset. The distribution is dominated by *stuff* classes. *Thing* classes constitute 1.51% of the pixel distribution.

Numbers only tell a partial story (about the statistical distribution) of the dataset. Fig. 4 shows representative samples for the person class from (a) ILSVRC dataset [43], (b) ADE20k dataset [51], and (c) AeroScapes dataset. A deep convolutional network trained on ILSVRC (source domain) is likely to not associate representations it learns for the person class with those for AeroScapes (target domain). However, ADE20k appears visually similar to AeroScapes for the person class. In Section 5, we observe that the visual appearance of object categories affects the performance of the system on the particular class.

4. Ensemble Knowledge Transfer

Our primary thesis is that the collective set of segmentation benchmarks represents a "meta" knowledge source that can be applied to a related, but different task. Importantly, each source encodes a considerable amount of curated human knowledge, manifested through the images and labels. We propose to *extract* this knowledge by training deep networks on each data source and *transfer* the knowledge to the target domain through fine-tuning.



(a) (b) (c)

Figure 4. Appearance of person class: **(a)** ILSVRC [43], **(b)** ADE20k [13], and **(c)** AeroScapes. While ILSVRC comprises of a million images representing a thousand classes, the visual appearance of the classes may be extremely different from scene parsing benchmarks such as ADE20k and AeroScapes. A network trained on ILSVRC will likely not associate the representations it learns for the person class with those in AeroScapes. However, ADE20k appears to be visually similar to AeroScapes.

The above procedure yields an ensemble of models, one for each data source, that can be applied to the target domain. Classic ensemble techniques may be used to aggregate the predictions, and compression techniques may distil the collective knowledge into a single network [17, 4]. We begin by discussing the intuitions which motivate us that this is a legitimate line of inquiry, particularly for the AeroScapes semantic segmentation setting.

4.1. Motivation

Symmetry and structure in natural scenes often results in unexpected visual correspondences. We qualitatively inspected the source domains [13, 36, 51] and target domain (AeroScapes) to understand whether objects appear to be visually similar across domains.

We discover a few predictable similarities - a potted plant may resemble a tree in an outdoor scene, and traffic signs and traffic lights may appear similar to obstacles such as streetlights (Fig. 5a). However, similarity in visual structure and symmetry may also occur in the absence of semantic similarity - a fan from an indoor scene may resemble an outdoor aerial drone, while a shower in an indoor scene may resemble a distant traffic light (Fig. 5b). Since we only transfer task agnostic knowledge from these source domains, such qualitiative similarities may translate into improvements in quantitative performance.

4.2. Data-driven Knowledge Transfer

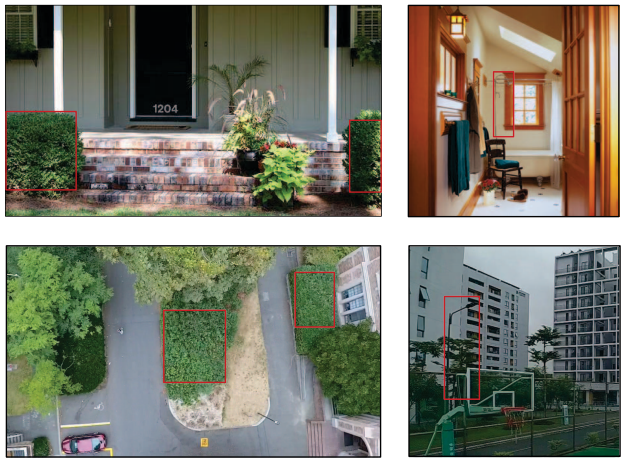
Knowledge transfer relies on preserving knowledge acquired while learning one task and applying it to another task. The simultaneous availability of a large amount of pixel annotations for specific domains and the *catastrophically forgetful* [35] nature of deep networks motivates us to study knowledge transfer in a data-driven manner as a means for solving tasks where limited amount of annotations are available. Specifically, we propose the transfer of knowledge from visually diverse domains to learn improved predictions for target domains with limited data.

In the supervised learning setting, we have a set of source domains, $D_s, \forall s \in \{1, 2, ..., S\}$, whose knowledge is compactly represented in the corresponding set of classifiers, $C_s, \forall s \in \{1, 2, ..., S\}$, which can be adapted for the task in the target domain D_{target}. Let X_{target} be an image in D_{target}, and Y_{target} be its associated label. We use the projection of X_{target} in domain D_s via classifier C_s to obtain the representation P_s. This helps us incorporate the knowledge from domain D_s:

$$C_s(X_{target}) \Rightarrow P_s$$

The complementary information encoded in each of the representations, P_s, is further used to learn a function, f, which aggregates them to predict the target domain label $\hat{Y_{target}}$:

$$f(P_1, P_2, ..., P_S) \Rightarrow \hat{Y_{target}}$$



(a) (b)

Figure 5. Similarity in visual structure and symmetry may occur in absence of semantic similarity - **(a)** a potted plant in PASCAL VOC appears visually similar to vegetation in Aeroscapes, and **(b)** a shower in ADE20k appears similar to an outdoor streetlight in AeroScapes. While potted plants are expected to be visually similar to vegetation, it is surprising to observe structural similarity between indoor showers and outdoor street lights. Since we only transfer class agnostic knowledge, qualitiative similarities may translate into improvements in quantitative performance.

4.3. Transferring Representations Across Domains

State-of-art semantic segmentation methods are based on deep neural networks. Our pixel classifiers, C_s, take the form of Fully Convolutional Networks (FCNs). A number of architectures have recently been proposed [8, 27, 28]. However, we choose to use the simple and effective vanilla FCN architecture for our analysis.

Since neural networks consist of millions of parameters and are quite sensitive to training data distributions, it is not wise to directly use them as feature extractors for the target domain. We adapt the projections P_s from domain D_s to the target domain by finetuning the higher task-specific layers of the FCNs while freezing the lower layers. We believe that finetuning the networks partially is the correct strategy for the following reasons: (1) Finetuning a smaller number of parameters in the network avoids overfitting for target domains with limited data. (2) *Critically*, finetuning all layers may result in the loss of complementary information that exists in different source domains. Finetuning only the task-specific layers enables the ensemble of networks to leverage knowledge from the diverse source domains.

4.4. Learning Representation Ensembles

We intend to learn an optimal method for combining the representations(P_s) produced by the classifiers(C_s). Formally, we seek to learn a function $f(C_1, C_2, ..., C_S; \theta)$, which predicts the segmentation label at each pixel location.

Inspired by the hypercolumn formulation [16], we combine the S model predictions by concatenating the class-probability distribution at each spatial location. Given a training image, X_i, and its ground truth segmentation map, Y^i_{target}, we seek to optimize the following objective:

$$\min_{\theta} \sum_{i} \|f(P^i_1, P^i_2, ..., P^i_S); \theta) - Y^i_{target}\|^2$$

We model $f(\cdot; \theta)$ as a single-layer regression network to learn the degree of contribution of each independent source domain for each class. In Sec. 5.2, we compare this regression network to other strategies for combining the predictions from each source domain.

5. Experimental Analysis

In this section, we explore the proposed ensemble knowledge transfer method for improving the performance of semantic segmentation tasks. The analysis is performed using the Cityscapes [11], PASCAL Context [36], and ADE20k [51] scene parsing segmentation benchmarks serving as the *source* domains and the AeroScapes dataset (Section 3) serving as the *target* domain.

We begin with a brief description of the methodology we follow for learning models for the AeroScapes dataset on the independent source domains, and the ensemble knowledge transfer network design for combining these single-source models. These descriptions are accompanied by analyses for the performance of these models. We conclude with analysis which demonstrates that complementary information from diverse source domains improves the performance of the multi-source ensemble.

Implementation Details: We use Fully Convolutional Networks [30](FCNs) for all experiments. We train the deep networks (Sec. 5.1) via Stochastic Gradient Descent using a minibatch size of one, $1e{-}10$ fixed learning rate, 0.99 momentum, and $5e-4$ weight decay. For each source domain, we freeze the first nine convolutional layers of the network and finetune the successive layers. The AeroScapes Dataset is divided into a $80\% - 20\%$ train-test splite. We ensure that image frames from a video sequence are only included in either training or testing. Throughout our experiments, the mean Intersection Over Union (mIOU) metric is used to report segmentation performance. The regression networks (Sec. 5.2) are trained with fixed $1e-2$ learning rate, 0.9 momentum, and $5e-4$ weight decay. The Caffe toolbox [20] is used to implement the networks.

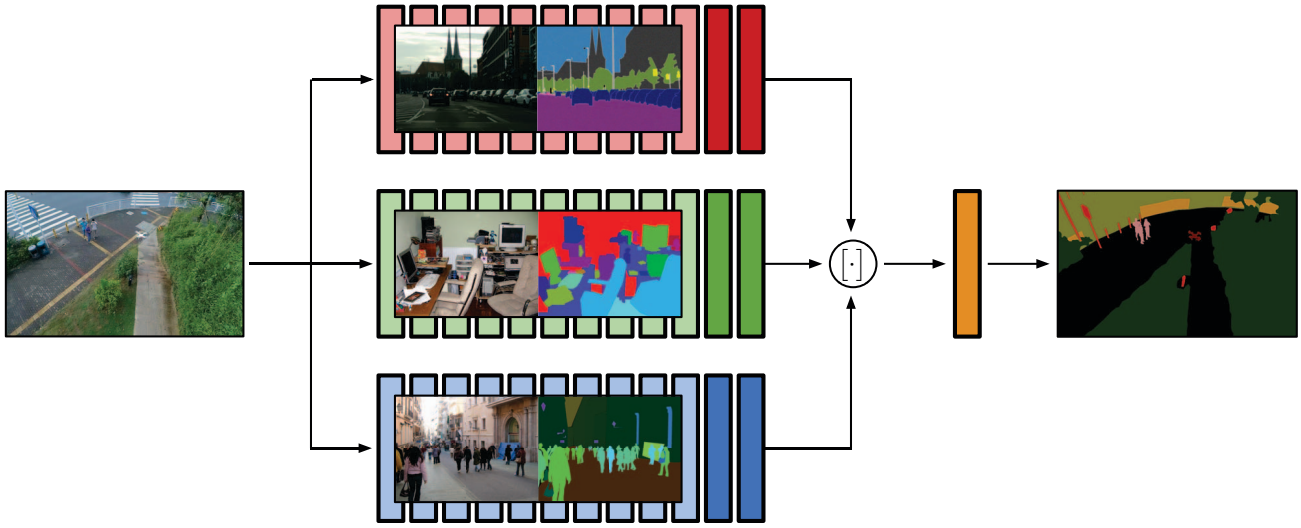


Figure 6. The AeroScapes dataset is used to finetune the higher (task-specific) layers of convolutional networks trained on independent source domains. Lower layers in the networks are not modified to preserve complementary information derived from diverse source domains. Finetuned representations are concatenated and a regressor is learned on the combined representations for the final prediction.

5.1. Learning from Single Sources

Recently, the practice of finetuning FCN-style networks on PASCAL VOC dataset has taken the intermediate step of finetuning on the MS COCO dataset [29]. This has resulted in small, but non-trivial, performance improvements [50]. Similarly, we finetune several FCN 8-stride networks pre-trained from public segmentation benchmarks towards the prediction on the Aerospace dataset. The source domains we use are PASCAL Context [36], ADE20k [51], and Cityscapes [11]. Note the PASCAL Context and PASCAL VOC [13] datasets contain overlapping images, but with distinct segmentation maps.

We perform an empirical analysis of the proposed framework. We first fine-tune the VGG-16 convolutional network [44] pre-trained on Imagenet (ILSVRC) [43] towards AeroScapes as a baseline method. We finetune VGG-16 networks that are pre-trained on ILSVRC to obtain an 8-stride FCN network. Since AeroScapes contains many small-scaled object categories, we also train 4-stride and 2-stride FCN networks. While we observe performance improvements on training the FCN 4-stride network over the FCN 8-stride network, the FCN 2-stride network does not provide any significant improvements over the FCN 4-stride network. We then repeat this procedure on pre-trained models from various domains, including PASCAL Context, ADE20k, and Cityscapes. For each source, we search over hyperparameters to find the best settings for knowledge transfer. This produces three different AeroScape models, that produce mean-IoUs of 52.02%, 51.62%, and 49.55%, respectively. The class-wise performance for each of these methods is detailed in Fig. 7.

Analysis A finer-resolution network trained on ILSVRC (FCN 4-stride) performs better than a coarser network (FCN 8-stride), except for for people and bicycles. We posit that a certain degree of "blurring" by operating at coarser resolutions helps knowledge transfer for such classes. This is likely to aid prediction since these classes are some of the most deformable *thing* classes in the AeroScapes dataseti - fine details may hurt predictions. FCN 8-stride networks initialized with other knowledge sources - PASCAL Context, Cityscapes, ADE20k - consistently out-perform FCN networks initialized and trained from ILSVRC data.

Predictably, AeroScapes models finetuned on certain domains do relatively better or worse on specific classes. Humans are of considerable interest in any segmentation benchmark. While PASCAL humans are primarily large foreground objects and Cityscapes humans are upright pedestrians or drivers, a non-trivial fraction of ADE20k humans (as illustrated in Fig. 4) are visually similar to AeroScapes humans. A model finetuned from Cityscapes performs better for construction but does worse on boats. Cityscapes consists of several classes which are visually similar to construction in Aeroscapes, while there are no boats in Cityscapes. Surprisingly, the model derived from Cityscapes does worse on Aeroscapes cars. We believe this is due to the drastic visual difference of Cityscapes cars which consist of front and rear view images as opposed to Aeroscapes which are primaly top-view car images. This inhomogenity in class-wise performance motivates us to combine the predictions from models finetuned on different source domains.

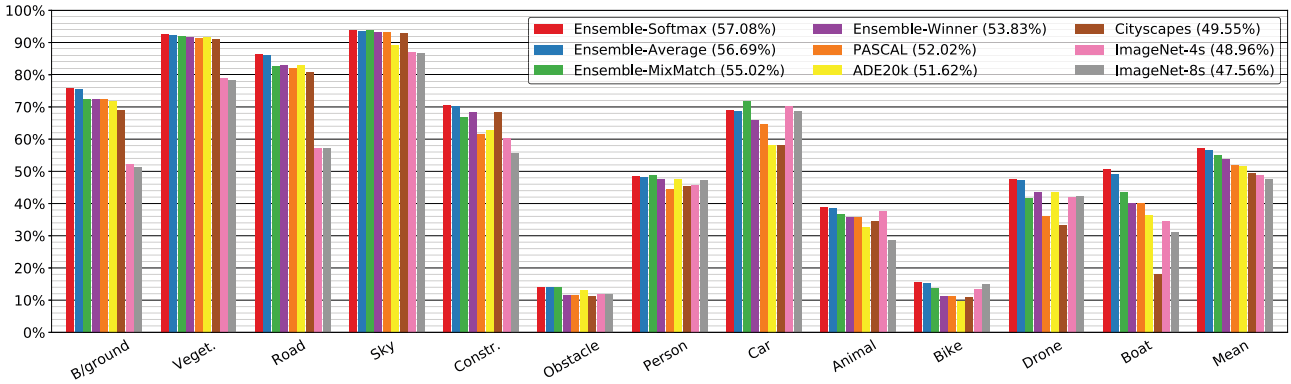


Figure 7. Comparison of methods. FCN 4-stride (**Imagenet-4s**) and FCN 8-stride (**ImageNet-8s**) networks are trained (initialized on ILSVRC) as baseline methods. Single-source single-network models are trained initialized from PASCAL Context (**PASCAL**), Cityscapes (**Cityscapes**), and ADE20k (**ADE20k**). Ensembles are created by several different strategies: winner-takes-all **Ensemble-Winner**, mix-and-match **Ensemble-MixMatch**, average ensemble **Ensemble-Average**, and weighted average ensemble **Ensemble-SoftReg** (refer to text for details). All models are FCN 8-stride networks, unless otherwise mentioned. The legend indicates mean IOU for each method.

5.2. Learning from Multiple Sources

Since certain pre-trained models do better on specific classes, it is natural to explore a *winner-take-all* approach: for each class, select the best single-source model. This strategy produces 53.83% mIOU (**Ensemble-Winner** in Fig. 7), which is 1.8% better than the best single-source model. While this suggests that combining sources is helpful, this is not a realizeable model.

The above strategy may be realized as a tangible system by combining the softmax distributions obtained from models that are learnt on single sources (Sec.5.1). We begin with a *mix-and-match* approach, assimilating softmax distributions from the single-source models based on class-wise winners. This model produces 55.02% mIOU (**Ensemble-MixMatch** in Fig. 7), which is 1.2% better than the *winner-take-all* approach.

The *mix-and-match* strategy provides an improvement over the *winner-takes-all* approach. However, it ignores all representations except class-winners and discards useful information. The simplest strategy to combine representations from each single-source model for each class is to average the softmax predictions. This *average ensemble* approach produces 56.69% mIOU (**Ensemble-Average** in Fig. 7), which is 1.6% better than the *mix-and-match* approach. This approach assumes that all softmax distributions are equally important for each class. Since we observe in Sec 5.1 that certain single-source models are relatively better or worse on specific classes, we now learn to weigh and combine the predictions from each source network. Specifically, we train a single layer regression network that learns to linearly combine the softmax distribution across the single-source models.

The proposed framework (Fig. 6), which is a *weighted average ensemble* of networks derived via late fusion of the softmax distributions produces 57.08% mIOU (**Ensemble-SoftReg** in Fig. 7), which is 0.4% better than the *average ensemble* approach. The regression network is trained with stratified sampling to ensure that the network is not biased towards *stuff* classes. We show qualitative results in Fig. 8.

Analysis: Limiting finetuning to the upper task-specific layers assists multi-source transfer as the ensembled models are diverse. **Ensemble-MixMatch** outperforming **Ensemble-Winner** suggests that it is better at handling negatives, which the IOU metric is sensitive to. **Ensemble-Average** outperforming **Ensemble-MixMatch** indicates that representations learnt from complementary domains are important for specific classes. **Ensemble-Average** performs surprisingly well, which indicates that the ensemble of networks learns quite potent complementary representations and simple aggregation works reasonably well. The sole category where we observe a non-trivial difference between **Ensemble-Average** and **Ensemble-SoftReg** is the boat class. This is likely due to the Cityscapes single-source model performing poorly on boats and degrading the **Ensemble-Average** boat classifier.

Single-source ensembles: We also investigate the source of the performance gains in the proposed framework - is the higher performance of the multi-source ensemble a function of complementary knowledge from *multiple sources* or simply a function of increased capacity due to *ensembling*? We train ensemble networks of equivalent capacity on *singular* source domains. Fig. 9 shows that single-source ensembling helps to an extent. However, single-source ensembles (53.05% mIOU) do not do as well as our proposed multi-source approach (57.08% mIOU).

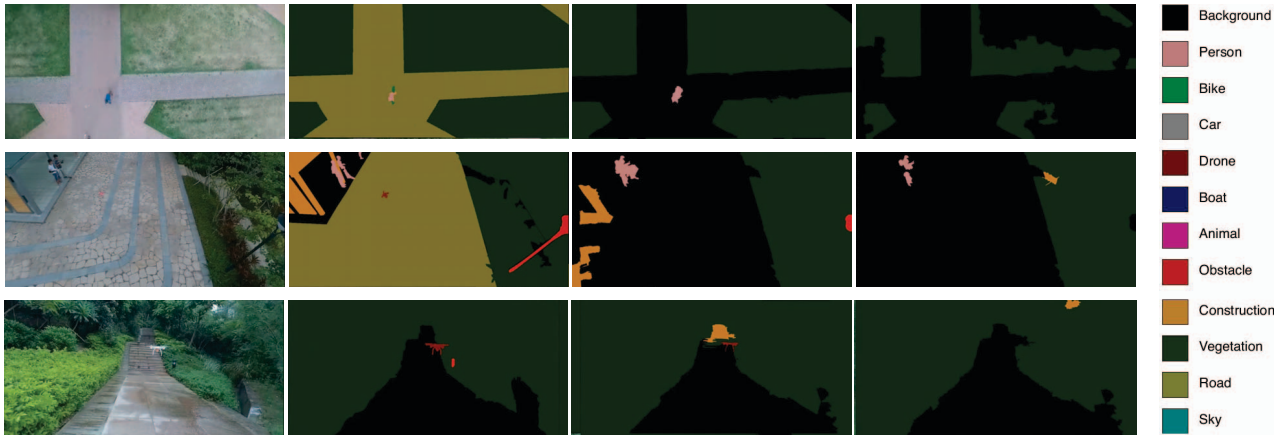


Figure 8. Each row shows image, ground-truth, proposed model (**Ensemble-SoftReg**), best single-source model (**PASCAL**). Row 1: proposed model segments human, but single-source model fails. Row 2: proposed model segments humans and also identifies obstacle partially, but single-source model does not. Row 3: single-source model does not detect the drone but proposed model segments it.

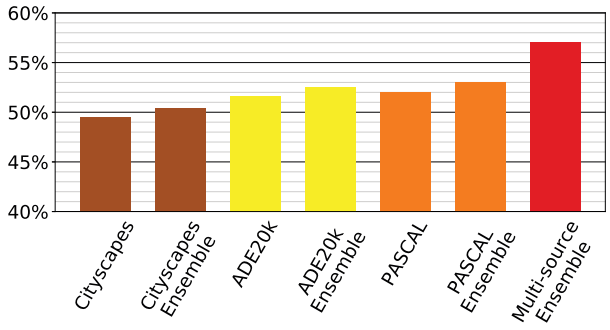


Figure 9. Comparisons of single-source models and multi-source models. The first, third, and fifth models represent performance when a single network is finetuned from a single source domain. The second, fourth, and sixth models represent performance when an ensemble of models is finetuned from a single source domain. The seventh model represents our proposed framework - an ensemble of models finetuned from diverse source domains. While we observe small performance improvements for single-source ensembles over their single-source single-network counterparts, the multi-source ensemble substantially supersedes the other methods.

6. Conclusion

Fully Convolutional Networks (FCNs) have established state-of-the-art performance on existing semantic segmentation benchmarks. Data-driven methods trained in supervised settings usually suffer from performance drop in the presence of domain shifts. In this research, we explore FCNs for semantic segmentation across data distributions that dramatically differ in scene structure, viewpoints, and objects statistics. We consider semantic segmentation on images with aerial viewpoints and study the transferability of knowledge from ground-view segmentation benchmarks. To this end, we prepare and release the AeroScapes dataset - a collection of 3269 aerial images (and associated semantic segmentation maps) captured using a fleet of drones.

We train multiple models for aerial segmentation via progressive fine-tuning from multiple source domains. The precise knowledge to be transferred from each domain is distinct and subtle - indoor objects can appear outdoors and outdoor objects may appear to be similar under aerial viewpoints. Thus, we treat the models tuned from different domains as an ensemble and aggregate them to improve performance. We successfully learn important components from each source domain through a regression network, resulting in an overall improvement of 8.12%.

The proposed framework is agnostic of the underlying network architecture and allows us to leverage small segmentation datasets that may comprise of critical complementary information. As future work, the network finetuning and prediction regression may be collaboratively learned to leverage information from diverse data sources.

Acknowledgments

This research would not have been possible without Autel Robotics, who helped us compile the AeroScapes dataset. This research was supported in part by the National Science Foundation (NSF) under grant IIS-1618903, the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001117C0051, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R & D Contract No. D17PC00345. Additional support was provided by Google Research and the Intel Science and Technology Center for Visual Cloud Systems (ISTC-VCS). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016. 1

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 1

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision*, 2016. 3

[4] C. Bucilu, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 4

[5] H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In *arXiv:cs-CV/1612.03716*, 2016. 3

[6] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998. 2

[7] A. Chaudhry, P. K. Dokania, and P. H. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *Proceedings of the British Machine Vision Conference*, 2017. 3

[8] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017. 1, 5

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv:cs-CV/170i7.05821*, 2017. 2

[10] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans. *Nature Scientific Reports*, 2016. 1

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3, 5, 6

[12] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[13] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012), 2010. 1, 4, 6

[14] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*, 2015. 3

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 3

[16] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[17] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Workshop on Deep Learning and Representation Learning*, 2014. 4

[18] G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 2

[19] J. Hoffman, D. Wang, F. Yu, and T. Darrell. FCNs in the Wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:cs-CV/1612.02649*, 2016. 3

[20] 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 *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014. 5

[21] S. Joon Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. In *Proceedings of the IEEE Conferene on Computer Vision and Pattern Recognition*, 2017. 3

[22] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, 2017. 1

[23] R. Kemker and C. Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. In *arXiv:cs-CV/1703.06452*, 2017. 1, 2

[24] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[25] I. Kokkinos. Ubernet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[26] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 3

[27] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5

[28] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 1, 6

[30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 5

[31] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2015. 3

[32] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, 2016. 3

[33] Y. Long, Y. Gong, Z. Xiao, and Q. Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017. 1

[34] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[35] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109 – 165. 1989. 4

[36] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. 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*, 2014. 1, 2, 3, 4, 5, 6

[37] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[38] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 2

[39] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 3

[40] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv:cs-CV/1606.02147*, 2016. 1

[41] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, 2016. 2

[42] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 3

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 4, 6

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 6

[45] K. Sirinukunwattana, S. E. A. Raza, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016. 1

[46] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. MultiNet: Real-time joint semantic reasoning for autonomous driving. *arXiv:cs-CV/1612.07695*, 2016. 1

[47] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[48] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[49] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 2014. 2

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 6

[51] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 4, 5, 6