This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

On Symbiosis of Attribute Prediction and Semantic Segmentation

Mahdi M. Kalayeh, *Member, IEEE,* and Mubarak Shah, *Fellow, IEEE*

Abstract—Attributes are semantically meaningful characteristics whose applicability widely crosses category boundaries. They are particularly important in describing and recognizing concepts for which no explicit training example is given, *e.g., zero-shot learning*. Additionally, since attributes are human describable, they can be used for efficient human-computer interaction. In this paper, we propose to employ semantic segmentation to improve person-related attribute prediction. The core idea lies in the fact that many attributes describe local properties. In other words, the probability of an attribute to appear in an image is far from being uniform in the spatial domain. We build our attribute prediction model jointly with a deep semantic segmentation network. This harnesses the localization cues learned by the semantic segmentation to guide the attention of the attribute prediction to the regions where different attributes naturally show up. As a result of this approach, in addition to prediction, we are able to localize the attributes despite merely having access to image-level labels (weak supervision) during training. We first propose semantic segmentation-based pooling and gating, respectively denoted as SSP and SSG. In the former, the estimated segmentation masks are used to pool the final activations of the attribute prediction network, from multiple semantically homogeneous regions. This is in contrast to global average pooling which is agnostic with respect to where in the spatial domain activations occur. In SSG, the same idea is applied to the intermediate layers of the network. Specifically, we create multiple copies of the internal activations. In each copy, only values that fall within a certain semantic region are preserved while outside of that, activations are suppressed. This mechanism allows us to prevent pooling operation from blending activations that are associated with semantically different regions. SSP and SSG, while effective, impose heavy memory utilization since each channel of the activations is pooled/gated with *all* the semantic segmentation masks. To circumvent this, we propose Symbiotic Augmentation (SA), where we *learn* only one mask per activation channel. SA allows the model to either pick one, or combine (weighted superposition) multiple semantic maps, in order to generate the proper mask for each channel. SA simultaneously applies the same mechanism to the reverse problem by leveraging output logits of attribute prediction to guide the semantic segmentation task. We evaluate our proposed methods for facial attributes on CelebA and LFWA datasets, while benchmarking WIDER Attribute and Berkeley Attributes of People for whole body attributes. Our proposed methods achieve superior results compared to the previous works. Furthermore, we show that in the reverse problem, semantic face parsing significantly improves when its associated task is jointly learned, through our proposed Symbiotic Augmentation, with facial attribute prediction. We confirm that when few training instances are available, indeed image-level facial attribute labels can serve as an effective source of weak supervision to improve semantic face parsing. That reaffirms the need to jointly model these two interconnected tasks.

Index Terms—Attribute Prediction, Semantic Segmentation, Semantic Gating, Facial Attributes, Person Attributes

1 Introduction

Nowadays, state-of-the-art computer vision techniques allow us to teach machines different classes of objects, actions, scenes, and even fine-grained categories. However, to learn a certain notion, we usually need positive and negative examples from the concept of interest. This creates a set of challenges as the instances of different concepts are not equally easy to collect. Also, the number of learnable concepts is linearly capped by the cardinality of the training data. Therefore, being able to robustly learn a set of *sharable concepts* that go beyond rigid category boundaries is of tremendous importance. Visual attributes are one particular type of these *sharable concepts*. They are human describable and machine detectable. We can use attributes to describe a variety of objects, scenes, actions, and events. For example, we associate a person who is lying on a beach with the attribute *relaxed* or a cat that is chasing after a wool ball with the attribute *playing*.

- *M. M. Kalayeh and M. Shah are with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, 32816.*
E-mails: mahdi@eecs.ucf.edu, shah@crcv.ucf.edu
This work updates and extends our previous work [1].

Attributes are different from category labels in three major aspects. **First**, category labels are agnostic with respect to the intra-class variations that exist among different instances of a single category. Such flat representation cannot distinguish between a *grumpy* cat and a *joyful* one as it only sees them as cats. **Second**, attributes go across category boundaries. Hence, they can be used to potentially describe an exponential number of object categories (via different combinations) even if the associated category has never been observed before (*e.g* zero-shot learning). **Third**, unlike category labels that can be effectively inferred from the object itself, humans heavily rely on the contextual cues for the attribute prediction. Take the examples shown in Figure 1. If we only consider the bounding box around the dog, one would not assign the attribute *catching* to it. Instead, *running* may even be a valid attribute. However, leveraging contextual layout where the dog is floating in air, and close to a frisbee, provides human with sufficient indications to not only rule out the attribute *running* but also confidently label the dog with the attribute *catching*. Similarly, the table, food and plate, collectively serve as the context, building the ground for associating attribute *eating* to the person.

Considering the aforementioned characteristics of at-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 1: Examples of how contextual layout assists attribute prediction in wild. The *person* (on left) and the *dog* (on right) should be respectively labeled with the attributes *eating* and *catching*. This is hard to agree upon if we would have taken these object instances in isolation, out of their contexts *i.e* food and frisbee.

tributes, we hypothesize that the attribute prediction task would benefit from contextual cues if they are properly represented. One can organize the context supervision into three levels: image-level, instance-level and pixel-level. Image-level supervision represents the context as a binary vector indicating whether an instance of a certain category appears somewhere in the context. Therefore, it is blind to the spatial relationships that exist between underlying components *i.e* object instances in the scene. In the instance-level supervision, context is available in terms of a set of category label and bounding box tuples. That is, unlike the image-level, instance-level context supervision can model the spatial relationships in the scene. Lastly, in the pixel-level context supervision, we have access to the category labels in a per-pixel fashion. Obviously, this provides a much stronger supervision signal compared to the other two alternatives. In this work, we propose augmenting attribute prediction by transferring weakly pixel-level context supervision, from an auxiliary semantic segmentation task.

So far, we've explained attributes in general when they describe an instance of an object in a scene. However, the same is valid when attributes characterize variations of a certain object category. In this paper, we are interested in person-related, specifically facial and full body attributes. We view the concept of contextual cues, previously detailed for attributes of objects in the scene, as the natural correspondence of object attributes to the object parts and their associated layout in the spatial domain of the object boundary.

Naturally, attributes are "additive" to the objects (*e.g.*, glasses for person). It means that an instance of an object may or may not take a certain attribute, while in either case the category label is preserved (*e.g.*, a person with or without glasses is still labeled as person). Hence, attributes are especially useful in problems that aim at modeling intra-category variations such as fine-grained classification. Despite their additive character, attributes do not appear in arbitrary regions of the objects (*e.g.*, hat if appears, is highly likely to show up on the top of person's head). This notion is the basis of our work. *We hypothesize that the attribute prediction can benefit from localization cues*. Specifically, to detect an attribute, instead of processing the entire spatial domain at once, one should focus on the region in which that attribute naturally shows up. However, not all attributes have precise correspondences. For example, it is ambiguous from where in the face, we as humans, infer if a person is *young* or *attractive*. Hence, instead of hard-coding the correspondences, even where those seem clear (*e.g.* glasses with nose and eyes), we allow the model to *learn* how to leverage the localization cues that are transferred from a relevant auxiliary task to the attribute prediction problem.

Using bounding boxes to show the boundary limits of objects is a common practice in computer vision. However, regions that different attributes are associated to drastically vary in terms of appearance. For example, in a face image, one cannot effectively put a bounding box around the region associated to "hair". In fact, the shape of the region can be used as an indicative signal on the attribute. On top of that, we have the partial occlusion of object parts which introduces further challenges by arbitrarily deforming visible regions. Therefore, we need an auxiliary task that learns detailed pixel-wise localization information without restricting the corresponding regions to be of certain pre-defined shapes. Semantic segmentation has all the aforementioned characteristics. It is the problem of assigning class labels to every pixel in an image. As a result, a successful semantic segmentation approach has to learn pixel-level localization cues which implicitly encode color, structure, and geometric characteristics in fine detail. In this work, since we are interested in person-related attributes, we take face [2] and human body [3] semantic parsing problems as auxiliary tasks to steer the spatial focus of the attribute prediction methods accordingly.

To perform attribute prediction, we feed an image to a fully convolutional neural network which generates feature maps that are ready to be aggregated and passed to the classifier. However, global pooling [4] is agnostic to where, in spatial domain, the attribute-discriminative activations occur. Hence, instead of propagating the attribute signal to the entire spatial domain, we funnel it into the semantic regions. By doing so, our model learns *where* to attend and *how* to aggregate the feature map activations. We refer to this approach as Semantic Segmentation-based Pooling (SSP), where activations at the end of the attribute prediction pipeline are pooled within different semantic regions.

Alternatively, we can incorporate the semantic regions into earlier layers of the attribute prediction network with a gating mechanism. Specifically, we propose augmenting max pooling operations such that they do not mix activations that reside in different semantic regions. Our approach generates multiple versions of the activation maps that are masked differently and presumably discriminative for various attributes. We refer to this approach as Semantic Segmentation-based Gating (SSG).

Since the semantic regions are not available for the attribute benchmarks, we learn to *estimate* them using a deep semantic segmentation network. In our earlier work [1], we took a conceptually similar approach to [5] in which an encoder-decoder model was built using convolution and deconvolution layers. However, considering the relatively small number of available data for the auxiliary segmentation task, we had to modify the network architecture. Despite being much simpler than [5], we found our semantic segmentation network [1] to be very effective in solving the auxiliary task of semantic face parsing. Examples of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

segmentation masks generated for previously unseen images are illustrated in Figure 2. Once trained, such network was able to provide localization cues in the form of masks (decoder output) that decompose the spatial domain of an image into mutually exclusive semantic regions. We show that both SSP and SSG mechanisms outperform almost all the existing state-of-the-art facial attribute prediction techniques while employing them together results in further improvements.

One issue with SSP and SSG is their memory utilization. Since both architectures use the output of semantic segmentation to create N_S (referring to the number of semantic regions) copies of the previous convolution layer activations. Given limited GPU memory budget, this can restrict the application of these layers when N_S grows to large values. Instead, we can circumvent this challenge by learning the proper mask per channel. In contrast to SSP and SSG which mask each and every channel of activations with *all* the N_S semantic probability maps, in this paper we propose to learn one mask per channel, as weighted superposition of different semantic probability maps (output of semantic segmentation network). Such workaround that can be simply implemented by a 1×1 convolution, adds minimum memory utilization overhead and also allows us to simplify SSP and SSG, yielding a single unified architecture that based on where it is applied in the architecture, can mimic the behavior of SSP and SSG.

Following the recent trend in semantic segmentation, instead of an encoder-decoder as in [1], here we utilize a fully convolutional architecture, specifically Inception-V3 [6]. Hence, we can unify attribute prediction and semantic segmentation networks by full weight sharing. As a result, unlike [1], we do not need to pre-train the semantic segmentation network before deploying it in attribute prediction pipeline. Instead, both tasks are learned simultaneously in an end-to-end fashion within a single architecture. We go beyond facial attributes [1] and demonstrate the effectiveness of employing semantic segmentation in person-related attributes on multiple benchmarks. Finally, we provide comprehensive quantitative evaluation for the case where attributes are jointly trained with semantic segmentation with the aim to boost the latter task.

In summary, the contributions of this work are as follows:

- We demonstrate the effectiveness of employing semantic segmentation to improve person-related attribute prediction.
- We propose a simple alternative to Semantic Segmentation-based Pooling and Semantic Segmentation-based Gating with focus on minimum memory utilization overhead.
- We unify semantic segmentation and attribute prediction through multi-tasking a single network and training it in an end-to-end fashion.
- We achieve state-of-the-art results in person-related attribute prediction on CelebA, LFWA, WIDER Attributes, and Berkeley Attributes of People datasets.
- We provide comprehensive experiments, detailing how to improve semantic segmentation task by leveraging image-level attribute annotations.

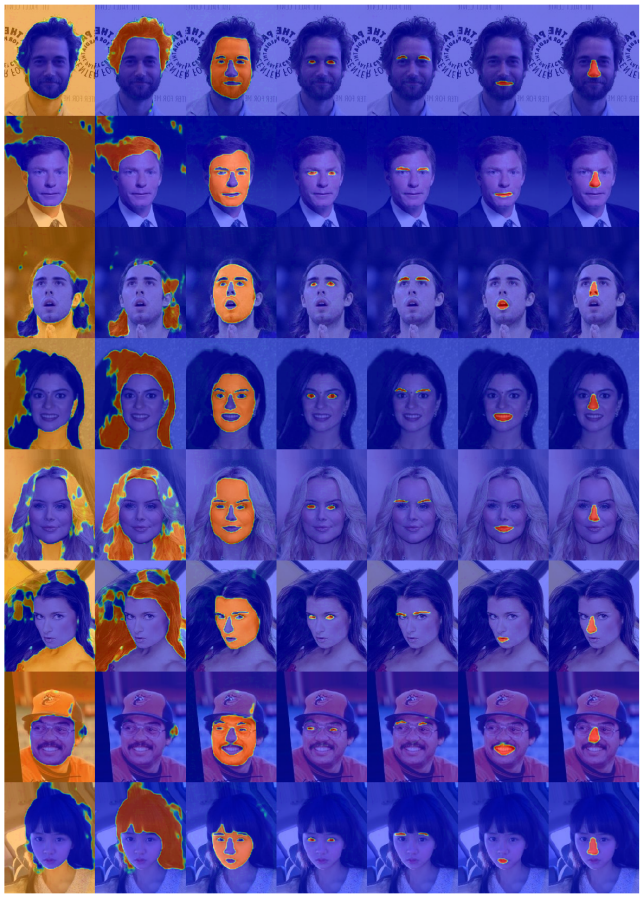


Fig. 2: Examples of the segmentation masks generated by our semantic segmentation network [1] for previously unseen images. From left to right: background, hair, face skin, eyes, eyebrows, mouth and nose.

The remainder of this paper is organized as follows. Section 2 offers a detailed review of attribute prediction and semantic segmentation literature. In Section 3, we propose semantic segmentation-based pooling and gating, followed by a simple unifying view of them which benefits from considerably lighter memory footprint. We end this section by providing details of our architectures. Experimental results are shown in Section 4. This includes evaluation of facial and person attributes on four datasets, alongside with comprehensive experiments on the effectiveness of leveraging image-level facial attribute annotations to boost semantic face parsing. Finally, we conclude the paper in Section 5.

2 Related Work

2.1 Attribute Prediction

Early works in modeling attributes [7] [8] [9] came around with the intention to change the recognition paradigm from naming objects to describing them. Therefore, instead of directly learning the object categories, one begins with learning a set of attributes that are shared among different categories. Object recognition can then be built upon the attribute scores. Hence, novel categories are seamlessly integrated, via attributes, with previously observed ones. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

can be used to ameliorate label misalignment between train and test data.

Considering the importance of human category, research in person-related attribute prediction [10] [11] [12] [13] [14] [15] has flourished over the years. To perform attribute prediction, some of the previous works have invested in modeling the correlation among attributes [16] [17] [18] [19], while others have focused on leveraging the category information [20] [21] [22]. There are also efforts to exploit the context [23].

Another way to view the attribute prediction literature is to divide it into holistic versus part-based methods. The common theme among the holistic models is to take the entire spatial domain into account when extracting features from images. On the other hand, part-based methods begin with an attribute-related part detection and then use the located parts, in isolation from the rest of spatial domain, to extract features. It has been shown that part-based models generally outperform the holistic methods. However, they are prone to the localization error as it can affect the quality of the extracted features. Although, there are works that have taken a hybrid approach benefiting from both the holistic and part-based cues [24] [12]. Our proposed methods fall in between the two ends of the spectrum. While we process the image in a holistic fashion, we employ localization cues in form of pixel-level semantic representations.

Among earlier works we refer to [10] [14] [15] [25] as successful examples of part-based attribute prediction models. More recently, in an effort to combine part-based models with deep learning, Zhang *et al.* [25] proposed PANDA, a pose-normalized convolutional neural network (CNN) to infer human attributes from images. PANDA employs poselets [15] to localize body parts and then extracts CNN features from the located regions. These features are later used to train SVM classifiers for attribute prediction. Inspired by [25], while seeking to also leverage the holistic cues, Gkioxari *et al.* [24] proposed a unified framework that benefits from both holistic and part-based models through utilizing a deep version of poselets [15] as part detectors. Liu *et al.* [12] have taken a relatively different approach. They show that pre-training on massive number of object categories and then fine-tuning on image level attributes is sufficiently effective in localizing the entire face region. Such weakly supervised method provides them with a localized region where they perform facial attribute prediction. In another part-based approach, Singh *et al.* [26] use spatial transformer networks [27] to localize the most relevant region associated to a given attribute. They encode such localization cue in a Siamese architecture to perform localization and ranking for relative attributes. Rudd *et al.* [28] have addressed the widely recognized data imbalance issue in attribute prediction, by introducing mixed objective optimization network (MOON). The proposed loss function mixes multiple task objectives with domain adaptive re-weighting of propagated loss. [29] and [30] are more examples of recent works that have tried similarly to address the class imbalance in the multi-label problem of attribute prediction. Li *et al.* have recently proposed lAndmark Free Face AttrIbute pRediction (AFFAIR) [31], a hierarchy of spatial transformation networks that initially crop and align the face region from the entire —assumed to be in the wild —input image and then localize relevant parts associated with different attributes. Separate neural network architectures then extract feature representations from global and part-based regions where their fusion is used to predict different facial attributes.

In our earlier work [1], we proposed employing semantic segmentation to capture local characteristics for facial attribute prediction. We utilized semantic masks, obtained from a separate pre-trained semantic segmentation network, to gate and pool the activations, respectively at middle and the end of the attribute prediction architecture. In this journal version of the paper, we extend and improve the proposed framework in [1] beyond face, and to the human body within the context of person-related attribute prediction. Our driving force in obtaining local cues is semantic parsing of face and human body. Meanwhile, unlike [1] that uses two separate networks for the main and auxiliary tasks, here we employ a heavy weight sharing strategy, unifying the semantic segmentation and attribute prediction architectures into one. Next, we discuss the semantic segmentation literature.

2.2 Semantic Segmentation

Semantic segmentation can be seen as a dense pixel-level multi-class classification problem, where the spatial (spatio-temporal) domain of images (videos) is partitioned using fine contours (volumes) into clusters of pixels (voxels) with homogeneous class labels. Prior to the wide-spread popularity of deep convolutional neural networks (CNN), semantic segmentation used to be solved via traditional classifiers such as Support Vector Machine (SVM) or Random Forest applied to the super-pixels [32] [33]. Conditional Random Field (CRF) was often used in these methods as the post processing technique to smooth the segmentation results, based on the assumption that pixels which fall within a certain vicinity, with similar color intensity, tend to be associated with the same class labels.

Among earlier efforts in using deep convolutional neural networks for semantic segmentation, we can refer to Ciresan *et. al* [34] work on automatic segmentation of neuronal structures in electron microscopy images. Although, since the number of classes was limited to only membrane and non-membrane, their problem in fact reduces to foreground detection task. Later, upon tremendous success of deep convolutional neural networks in image classification, researchers began designing semantic segmentation models on the top of CNN models, which were previously trained for other tasks, mainly image classification [35] [36] [37] [38] [39]. These methods, by leveraging supervised pre-training on strongly correlated tasks (*e.g.* often labels in two tasks have some overlap), were able to facilitate training procedure for semantic segmentation. However, such an adoption introduces its very own challenges.

Unlike image classification where the activations just before the classifier are flattened via fully connected layer or global average pooling, semantic segmentation task requires the spatial domain to be maintained, specifically the output segmentation maps should be at least of the same size as the input image. Fully Convolutional Networks [35] popularized CNN architectures for semantic segmentation. Long

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

et. al [35] proposed transforming fully connected layers into convolution layers along with up-sampling intermediate and final activations, whose spatial domain have reduced due to pooling layers through the network architecture. These techniques enable a classification model to output segmentation maps of arbitrary size when operating on input images of any size. Almost all the subsequent state-of-the-art semantic segmentation methods adopted this paradigm. The performance of semantic segmentation task will be compromised if the spatial information is not well preserved through the network architecture. In contrast, architectures designed for image classification very often use pooling layers to aggregate the context activations while discarding the precise spatial coordinates. To alleviate this conceptual discrepancy, two different classes of architectures have evolved.

First is the encoder-decoder based approach [5] in which the encoder gradually reduces the spatial domain through successive convolution and pooling layers, to generate the bottleneck representation. Then the decoder recovers the spatial domain by applying multiple layers of deconvolution or convolution followed by up-sampling, to the aforementioned bottleneck representation. There are usually shortcut connections from the encoder to the decoder, leveraging details at multiple scales, in order to help decoder recovering fine characteristics more accurately. U-Net [40] SegNet [36], and RefineNet [41] are the popular architectures from this class.

The second class of architectures developed around the idea of Dilated or Atrous convolutions [37]. Specifically, one can avoid using pooling layers in order to preserve detailed spatial information, but this will dramatically increase the computation cost as the following layers must operate on larger activation maps. However, using Atrous convolution [37] with dilation rate equal to the stride of the avoided pooling layer, results in the exact same number of operations as the regular convolution operating on pooled activations[1]. In other words, dilated or Atrous convolution layer allows for an exponential increase in effective receptive field without reducing the spatial resolution. In a series of works [42] [38], Chen *et. al.* demonstrated how Atrous convolution and its multi-scale variation, namely Atrous spatial pyramid pooling (ASPP) can be utilized within the framework of fully convolutional neural networks to improve the performance of the semantic segmentation task. While in earlier efforts [38], Dense CRF [39] has been used, more recent works [42] have shown comparable results without using such post-processing technique.

Semantic segmentation can be applied at a finer granularity where instead of the entire scene, an object is semantically parsed into its parts. Among popular examples, readers are encouraged to refer to [2] [43] [44] [45] for face, [46] [47] [48] [49] [50] [51] for general objects, and [3] [52] [53] [54] [55] [56] [57] [58] [59] for human body and clothing semantic parsing.

In this work, since we are interested in attributes describing human, when alluding to semantic segmentation, we specifically mean face and human body semantic parsing. Our semantic segmentation model is a fully convolutional neural network based on Inception-V3 [6] architecture, where following [38] [42] we have also incorporated Atrous spatial pyramid pooling (ASPP). In addition to utilizing semantic parsing for person-related attribute prediction, we will provide results on semantic face parsing as well. We show that, training an attribute prediction network with image-level supervision can effectively serve as an initialization for semantic parsing task, when the the number of training instances is limited.

1. It is worth pointing out that while the computation cost remains the same, employing dilated convolution demands more memory since the size of activation maps remains intact.

3 Methodology

The underlying idea of this work is to exploit semantic segmentation in order to improve person-related attribute prediction. To do so, we first revisit semantic segmentation-based pooling (SSP) and gating (SSG), initially proposed in our earlier work [1]. Then, we propose a considerably simpler architecture, which unifies SSP and SSG designs while approximately mimicking their behavior with drastically smaller memory footprint. Furthermore, unlike [1], where there were two networks, one for semantic segmentation and the other for attribute prediction, here we unify two networks with fully sharing the weights among two tasks, and train in an end-to-end fashion. Note that in [1], once trained independently, the semantic segmentation network was frozen during the attribute prediction task. Moving towards more modern architectures than those used earlier in [1], we describe our new models based on modern Inception-V3 [6] as their backbone. This choice will allow us to further push performance boundaries in person-related attribute prediction task.

3.1 SSP: Semantic Segmentation-based Pooling

We argue that attributes usually have a natural correspondence to certain regions within the object boundary. Hence, aggregating the visual information from the entire spatial domain of an image would not capture this property. This is the case for the global average pooling used in our baseline as it is agnostic to where, in the spatial domain, activations occur. Instead of pooling from the entire activation map, we propose to first decompose the activations of the last convolution layer into different semantic regions and then aggregate only those that reside in the same region. Hence, rather than a single vector representation, we obtain multiple features, each representing only one semantic region. This approach has an interesting intuition behind it. In fact, SSP funnels the back-propagation of the label signals, via multiple paths, associated with different semantic regions, through the entire network. This is in contrast with global average pooling that rather equally affects different locations in the spatial domain. We later explore this by visualizing the activation maps of the final convolution layer.

We can simply concatenate the representations associated with different regions and pass it to the classifier; however, it is interesting to observe if attributes indeed prefer one semantic region to another. Also, whether what our model learns matches human expectation on what attribute corresponds to which region. To do so, we take a similar

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

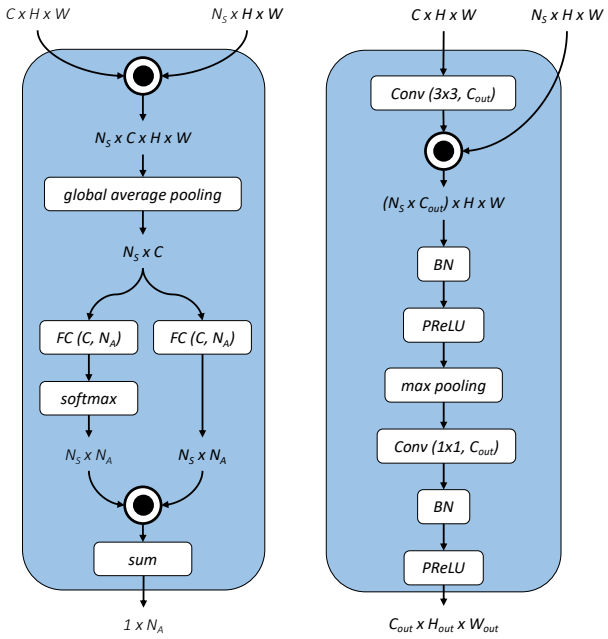


Fig. 3: Left: Semantic segmentation-based Pooling (SSP). Right: Semantic segmentation-based Gating (SSG). N_S and N_A, respectively, indicate the number of labels in semantic segmentation and attribute prediction tasks. We assume that the output tensor of activations from the previous layer to either SSP or SSG is of shape $C \times H \times W$ where C, H and W, respectively represent the number of channels, height and width of the activations. Alternatively, in Sec. 3.3, we will show that instead of using all N_S semantic regions for every channel, one can learn a single semantic mask per channel. This would also unify the SSP and SSG architectures.

approach to [60] where Bilen and Vedaldi employed a two branch network for weakly supervised object detection. We pass the vector representations, each associated with a different semantic region, to two branches one for recognition and another for localization. We implement these branches as linear classifiers that map vector representations to the number of attributes. Hence, we have multiple detection scores for an attribute each inferred based on one and only one semantic region. To combine these detection scores, we normalize outputs of the localization branch using softmax non-linearity across different semantic regions. This is a per-attribute operation, not an across-attribute one. We then compute the final attribute detection scores by a weighted sum of the per-region logits (*i.e.* outputs of recognition branch) using weights generated by the localization branch. Figure 3 (Left) shows the SSP architecture.

3.2 SSG: Semantic Segmentation-based Gating

Max pooling is used to compress the visual information in the activation maps of the convolution layers. Its efficacy has been proven in many computer vision tasks, such as image classification and object detection. However, attribute prediction is inherently different from image classification. In image classification, we want to aggregate the visual information across the entire spatial domain to come up with a single label for the image. In contrast, many attributes are inherently localized to specific image regions. Consequently, aggregating activations that reside in the "hair" region with the ones that correspond to "mouth", would confuse the model in detecting "smiling" and "wavy hair" attributes. We propose SSG to cope with this challenge.

Figure 3 (Right), shows our proposed SSG architecture C_{out} may or may not be the same as C (similarly for H and W). To gate the output activations of the convolution layer, we broadcast element-wise multiplication for each of the semantic regions with the entire activation maps. This generates multiple copies of the activations that are masked differently. In other words, such mechanism spatially decomposes the activations into copies, where large values cannot simultaneously occur in two semantically different regions. For example, gating with the semantic mask that corresponds to the "mouth" region, would suppress the activations falling outside its area while preserving those that reside inside it. However, the area which a semantic region occupies varies from one image to another.

We observed that, directly applying the output of the semantic segmentation network results in instabilities in the middle of the network. To alleviate this, prior to the gating procedure, we normalize the semantic masks such that the values of each channel sums up to 1. We then gate the activations right after the convolution and before the batch normalization [61]. This is very important since the batch normalization [61] enforces a normal distribution on the output of the gating procedure. Then, we can apply max pooling on these gated activation maps. Since, given a channel, activations can only occur within a single semantic region, max pooling operation cannot blend activation values that reside in different semantic regions. We later restore the number of channels using a 1×1 convolution. It is worth noting that SSG can potentially mimic the standard max pooling by learning a sparse set of weights for the 1×1 convolution. In a nutshell, semantic segmentation-based gating allows us to process the activations of convolution layers in a per-semantic region fashion while it also learns how to blend the pooled values back in.

3.3 A Simple Unified View to SSP and SSG

In both SSP and SSG architectures, we use the output of semantic segmentation to create N_S copies of the activations. Each copy, assuming semantic parsing outputs are perfect, preserves the activation values residing in one semantic region while suppressing those that are outside that. Hence, both SSP and SSG should maintain N_S times the size of activation maps in the memory. As N_S value grows, this can certainly become problematic due to limited GPU memory budget. A simple workaround for this is to learn the masks per channel. Specifically, instead of masking each and every channels of the previous convolution activations by *all* the N_S semantic probability maps, we learn one mask per channel (ref. Φ_S in Figure 4). This can be simply implemented via a 1×1 convolution on the top of semantic segmentation probability maps. However, in practice, we observed that larger kernels can result in slight performance gain. Similar to SSG, the output logits of the semantic segmentation classifier must be normalized, via

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

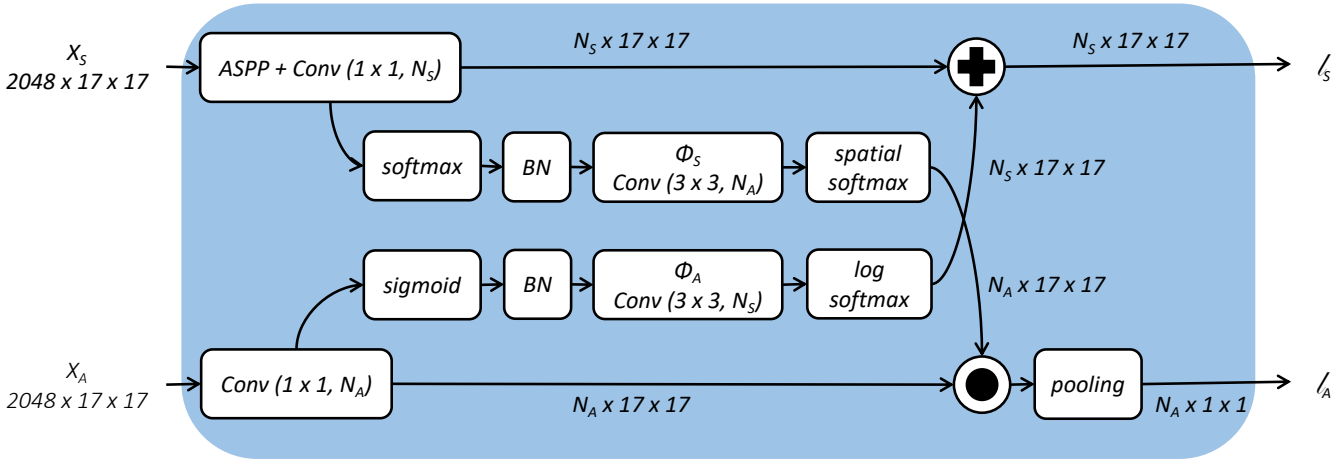


Fig. 4: Architecture of the Symbiotic Augmentation (SA). The embedding layers, Φ_S and Φ_A, respectively utilize the output of semantic segmentation and attribute prediction classifiers to augment the other task. Similar to Figure 3, N_S and N_A denote the number of output labels for semantic segmentation and attribute prediction, where, l_S and l_A are their corresponding loss functions (per-pixel softmax cross entropy, image-level sigmoid cross entropy). Addition and multiplication are element-wise operations.

batch normalization, prior to being passed to the embedding convolution layer. The output of the embedding should also be spatially normalized. Such embedding allows the model to either pick one or combine (weighted superposition) multiple semantic maps, in order to generate proper mask for each channel. We initialize the convolution kernels Φ_S of the embedding layers with zeros and no bias. This is inspired by the idea that each channel should start by using all the semantic regions equally. However, through training, it has the freedom to change towards combining only a selected number of regions. We later visualize how the learned convolution kernels of Φ_S look like in Figures 9 and 8a.

We now go one step further as the same idea can be used when we reverse the roles of tasks. In particular, we can use the output of attribute prediction to guide the semantic segmentation task. We refer to this joint semantic augmenting model, illustrated in Figure 4, as Symbiotic Augmentation (SA). The architecture of the embedding module in this case, Φ_A, is the same as Φ_S except the normalization operations are done differently. Figure 4 shows that in Symbiotic Augmentation, each task augments the other task's representation, through its corresponding output logits, while simultaneously being trained in an end-to-end fashion. This is different than SSP and SSG, where only a pre-trained semantic segmentation model, while frozen at deployment, augments attribute prediction task. Note that, in addition to a lower memory footprint[2], this approach allows us to simplify the SSP by unifying the recognition and localization branches. That is because the learned masks can properly weigh each channel and the order of consecutive linear operations (matrix multiplication through fully connected layer and scaling through weights of localization branch) is interchangeable.

2. The memory footprint of SSP is of $\mathcal{O}(N_S CHW)+\mathcal{O}(N_S N_A)$ while SA's is of $\mathcal{O}(N_S HW) + \mathcal{O}(N_A HW)$. Here C refers to the number of output channels in last (before classifier) convolution layer, while H and W respectively denote height and width of the final spatial resolution.

3.4 Network Architectures

We use Inception-V3 [6] as the convolutional backbone of Symbiotic Augmentation (SA), for both semantic segmentation and attribute prediction models. Its architecture is 48 layers deep and uses global average pooling instead of fully-connected layers which allows operating on arbitrary input image sizes. Inception-V3 [6] has a total output stride of 32. However, to maintain low computation cost and memory utilization, the size of activation maps quickly reduces by a factor of 8 in only first seven layers, referred to as STEM [6] in Figure 5. This is done by one convolution and two max pooling layers that operate with the stride of 2. The network follows by three blocks of Inception layers separated by two grid reduction modules. Spatial resolution of the activations remains intact within the Inception blocks, while grid reduction modules halve the activation size and increase the number of channels. For more details on the Inception-V3 [6] architecture, readers are encouraged to refer to [6]. Note that, for SSP, SSG and SSP+SSG experiments which were initially reported in [1], a VGG16-like backbone architecture has been used. Further details are provided in [1].

In this work, we use a single architecture to simultaneously learn semantic parsing and attribute prediction tasks. This is different than [1] where semantic segmentation model was pre-trained and then deployed (weights were frozen) into attribute prediction pipeline. Specifically, we share the weights of the Inception-V3 [6] while training with a mixed minibatch that is comprised of equal instances associated to attribute prediction and semantic segmentation tasks. Figure 5 illustrates how we obtain feature representations for both tasks using a single architecture. Note that each element in the minibatch has only one type of annotations, either attribute or semantic segmentation labels. Hence, when X_A and X_S are passed to the Symbiotic Augmentation (SA), shown in Figure 4, depending on the annotation type, either l_S or l_A are calculated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

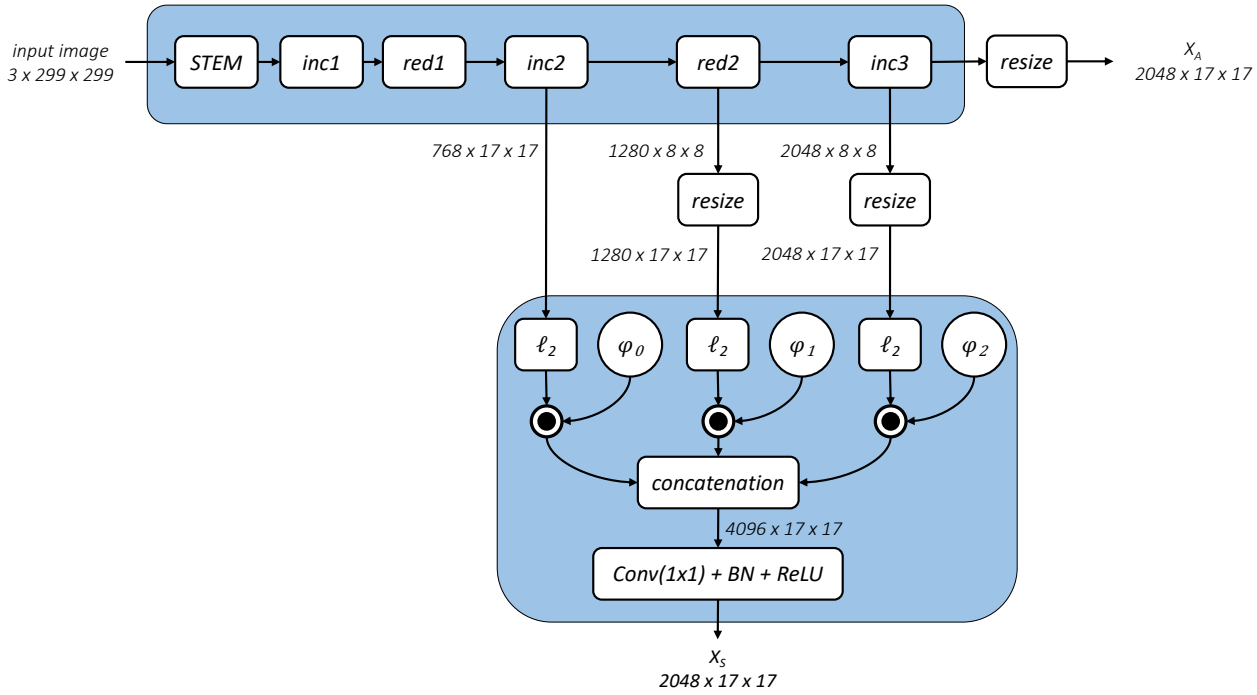


Fig. 5: Inception-V3 [6] backbone architecture used in the Symbiotic Augmentation (SA) experiments. X_A and X_S are used as input features to SA (ref. Figure 4). In order to generate X_S, we ℓ_2 normalize the intermediate activations and scale them by learnable φ_* parameters. Refer to [6] for the details of the Inception-V3 architecture.

4 Experiments

4.1 Datasets and Evaluation Measures

We evaluate our proposed attribute prediction models on multiple benchmarks. Specifically, we use CelebA and LFWA [12] for facial attributes, while benchmarking on WIDER Attribute [23] and Berkeley Attributes of People [15] for person attribute prediction.

Liu *et al.* [12] have used classification accuracy/error as the evaluation measure on CelebA and LFWA. However, we believe that due to significant imbalance between the numbers of positive and negatives instances per attribute, such measure cannot appropriately evaluate the quality of different methods. Similar point has been raised by [28], [29], [30] as well. Therefore, in addition to the classification error, we also report the average precision (AP) of the prediction scores. Following the literature, we solely report AP for WIDER Attribute [23] and Berkeley Attributes of People [15]. Since attribute benchmarks do not come with pixel-level labels, we train our semantic segmentation model on auxiliary datasets. For experiments corresponding to facial attributes, we use Helen Face [43] along with segment label annotations supplemented by [2]. For person attribute prediction experiments, we train the semantic parsing model on Look into Person (LIP) [3] dataset. We use the standard data split of each corresponding dataset.

CelebA [12] consists of 202,599 images partitioned into training, validation and test splits with approximately 162K, 20K and 20K images in the respective splits. There are a total of 10K identities (20 images per identity) with no identity overlap between evaluation splits. However, we do not use identity annotations. Images are annotated with 40 facial attributes such as, "wavy hair", "mouth slightly open", and "big lips". In addition to the original images, CelebA provides a set of pre-cropped images. We report our results on both of these image sets.

LFWA [12] has a total of 13,143 images of 5,749 identities with pre-defined train and test splits, which divide the entire dataset into two approximately equal partitions. Each image is annotated with the same 40 attributes used in CelebA [12].

WIDER Attribute [23] is collected from 13,789 WIDER images [62], containing usually many people in each image with huge human variations. Each person in these images is then annotated with a bounding box and 14 distinct human attributes such as "longhair", "sunglasses", "hat", "skirt", and "facemask". This results in a total of 57,524 boxes. Out of 13,789 images, WIDER Attribute [23] is split into 5,509 training, 1,362 validation and 6,918 test images. There are 30 scene-level labels that each image is annotated with. However, we do not use them and solely train and evaluate on bounding boxes of people. We evaluate on the 29,179 bounding boxes from testing images, after training on 28,345 person boxes extracted from aggregation of training and validation images. Unlike CelebA and LFWA [12], missing attributes are allowed in WIDER Attribute [23] dataset.

Berkeley Attributes of People [15] contains 4,013 training and 4,022 test instances. The example images are centered at the person and labeled with 9 attributes namely, "is male", "has long hair", "has glasses", "has hat", "has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

tshirt", "has long sleeves", "has shorts", "has jeans", "has long pants". Similar to the WIDER Attribute [23], here unspecified attributes are also allowed.

Helen Face [43] consists of 2,330 images with highly accurate and consistent annotations of the primary facial components. Smith *et. al* [2] have supplemented Helen Face [43] with 11 segment label [3] annotations per image. Images are divided into splits of 2000, 230 and 100, respectively for training, validation and test. We train our semantic segmentation model on the aggregation of training and validation splits and evaluate on the test split.

LIP [3] consists of ∼30,000 and 10,000 images respectively for train and validation. Each images is annotated with 20 semantic labels[4].

4.2 Evaluation of Facial Attribute Prediction

For all the numbers reported here, we want to point out that FaceTracer [11] and PANDA [25] use groundtruth landmark points to attain face parts. Wang *et al.* [63] use 5 million auxiliary image pairs, collected by the authors, to pre-train their model. Wang *et al.* [63] also use state-of-the-art face detection and alignment to extract the face region from CelebA and LFWA images. *However, we train all our models with only attribute and auxiliary face/human parsing labels*.

We compare our proposed method with the existing state-of-the-art attribute prediction techniques on the CelebA [12]. To prevent any confusion and have a fair comparison, Table 1 reports the performances in two separate columns distinguishing the experiments that are conducted on the original image set from those where the pre-cropped image set have been used.

Experimental results indicate that under different settings and evaluation protocols, our proposed semantic segmentation-based pooling and gating mechanisms can be effectively used to boost the facial attribute prediction performance. That is particularly important given that our global average pooling baselines already beat almost all the existing state-of-the-art methods. To see if SSP and SSG are complementary to each other, we also report their combination where the corresponding predictions are simply averaged. We observe that such process further boosts the performance.

To investigate the importance of aggregating features within the semantic regions, we replace the global average pooling in our basic model with the spatial pyramid pooling layer [65]. We use a pyramid of two levels and refer to this baseline as SPPNet*. While aggregating the output activations in different locations, SPPNet* does not align its pooling regions according to the semantic context that appears in the image. This is in direct contrast with the intuition behind our proposed methods. Experimental results shown in Table 1 confirm that simply pooling the output activations at multiple locations is not sufficient. In fact, it results in a lower performance than global average pooling. This verifies that the improvement obtained by our proposed models is due to their content aware pooling/gating mechanisms.

Classification Error(%)		
Method	Original	Pre-cropped
FaceTracer [11]	18.88	–
PANDA [25]	15.00	–
Liu *et al.* [12]	12.70	–
Wang *et al.* [63]	12.00	–
Zhong *et al.* [64]	10.20	–
Rudd *et al.* [28]: Separate	–	9.78
Rudd *et al.* [28]: MOON	–	9.06
AFFAIR [31]	8.55	–
SPPNet*	–	9.49
Naive Approach	9.62	9.13
BBox	–	8.76
Avg. Pooling	9.83	9.14
SSG	9.13	8.38
SSP	8.98	8.33
SSP + SSG	8.84	**8.20**
Inception-V3: baseline	8.68	–
Symbiotic Augmentation (SA)	**8.53**	–
Average Precision(%)		
Method	Original	Pre-cropped
AFFAIR [31]	79.63	–
SPPNet*	–	77.69
Naive Approach	76.29	79.74
BBox	–	79.95
Avg. Pooling	77.16	79.74
SSG	77.46	80.55
SSP	78.01	81.02
SSP + SSG	78.74	**81.45**
Inception-V3: baseline	79.28	–
Symbiotic Augmentation (SA)	**80.10**	–
Balanced Accuracy(%) [29]		
Method	Original	Pre-cropped
Huang *et al.* [29]	–	84.00
CRL(C) [30]	–	85.00
CRL(I) [30]	–	86.00
Avg. Pooling	–	86.73
SSG	–	87.82
SSP	–	**88.24**

TABLE 1: Attribute prediction performance evaluated by the classification error, average precision and balanced classification accuracy [29] on the CelebA [12] original and pre-cropped image sets.

Naive Approach A naive alternative approach is to consider the segmentation maps as additional input channels. To evaluate its effectiveness, we feed the average pooling basic model with 10 input channels, 3 for RGB colors and 7 for different semantic segmentation maps. The input is normalized using Batch Normalization [61]. We train the network using the same setting as other aforementioned models. Our experimental results indicate that such naive approach cannot leverage the localization cues as good as our proposed methods. Table 1 shows that at best, the naive approach is on par with the average pooling basic model. We emphasize that feeding semantic segmentation maps along with RGB color channels to a convolutional network results in blending the two modalities in an *additive* fashion. Instead, our proposed mechanisms take a *multiplicative*

3. "background", "face skin" (excluding ears and neck), "left eyebrow", "right eyebrow", "left eye", "right eye", "nose", "upper lip", "inner mouth", "lower lip" and "hair"

4. "Background", "Hat", "Hair", "Glove", "Sunglasses", "Upper-clothes", "Dress", "Coat", "Socks", "Pants", "Jumpsuits", "Scarf", "Skirt", "Face", "Right-arm", "Left-arm", "Right-leg", "Left-leg", "Right-shoe" and "Left-shoe"

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

Method	Classification Error(%)	AP(%)
FaceTracer [11]	26.00	–
PANDA [25]	19.00	–
Liu *et al.* [12]	16.00	–
Zhong *et al.* [64]	14.10	–
Wang *et al.* [63]	13.00	–
AFFAIR [31]	13.87	83.01
Avg. Pooling	14.73	82.69
SSG	13.87	83.49
SSP	13.20	84.53
SSP + SSG	**12.87**	**85.28**

TABLE 2: Attribute prediction performance evaluated by the classification error and the average precision (AP) on LFWA [12] dataset.

Method	AP(%)
Fast R-CNN [66]	80.00
R*CNN [67]	80.50
Deep Hierarchical Contexts [23]	81.30
VeSPA [68]	82.40
ResNet-101 [69]	85.00
ResNet-SRN-att [69]	85.40
ResNet-SRN [69]	86.20
Sarafianos *et. al.* [70]	86.40
Inception-V3: baseline	85.86
Symbiotic Augmentation (SA)	**87.58**

TABLE 3: Attribute prediction performance evaluated by the average precision(%) on WIDER Attribute [23] dataset.

Method	AP(%)
Fast R-CNN [66]	87.80
R*CNN [67]	89.20
Gkioxari *et al.* [24]	89.50
Deep Hierarchical Contexts [23]	92.20
Inception-V3: baseline	92.87
Symbiotic Augmentation (SA)	**94.80**

TABLE 4: Attribute prediction performance evaluated by the average precision(%) on Berkeley Attributes of People [15] dataset.

approach by masking the activations using the semantic segmentation probability maps.

Semantic Masks vs. Bounding Boxes To analyze the necessity of semantic segmentation, we generate a baseline, namely BBox, which is similar to SSP. However, we replace the semantic masks in SSP with the bounding boxes on the facial landmarks. Note that we use the groundtruth location of the facial landmarks, provided in CelebA dataset [12], to construct the bounding boxes. Hence, to some extent, the performance of BBox is the upper bound of the bounding box experiment. There are 5 facial landmarks including left eye, right eye, nose, left mouth and right mouth. We use boxes with area 20^2 (40^2 gives similar results) and 1:1, 1:2 and 2:1 aspect ratios. Thus, there are a total of 16 regions including the whole image itself. From Table 1, we see that our proposed models, regardless of the evaluation measure, outperform the bounding box alternative, suggesting that semantic masks should be favored over the bounding boxes on the facial landmarks.

Balanced Classification Accuracy Given the significant imbalance in the attribute classes, also noted by [28], [29], [30], we suggested using average precision instead of classification accuracy/error to evaluate attribute prediction. Instead, Huang *et al.* [29] and later [30] have adopted balanced accuracy measure. To evaluate our proposed approach in balanced accuracy measure, we fine-tuned our models with the weighted ($\propto$ imbalance level) binary cross entropy loss. From Table 1, we observe that under such measure, all the variations of our proposed model outperform both [29] and [30] with large margins.

To better understand the effectiveness of our proposed approach on facial attributes, we also report experimental results on the LFWA dataset [12] in Table 2. Here, we observe a similar trend to the one in CelebA, where all the proposed models which exploit localization cues successfully improve the baseline. Specifically, SSP + SSG achieves considerably better performance than the average pooling model with margins of 1.86% in classification accuracy and 2.59% in average precision. Our best model also outperforms all other state-of-the-art methods.

Symbiotic Augmentation (SA) All the results reported so far were using a VGG16-like architecture for attribute prediction and a separate pre-trained encoder-decoder architecture for semantic segmentation [1]. However, in SA-based models, we have unified the two architectures and train simultaneously with two objective functions. Table 1 shows that simply using a stronger convolutional backbone like Inception-V3 [6] boosts the performance on CelebA original image set. Furthermore, SA-based model which is built on the top of such backbone, despite heavily sharing all the weight across two tasks, can achieve even better results, outperforming SSP+SSG and current state-of-the-art AFFAIR [31]. However, on LFWA dataset [12], we observed that Inception-V3 [6] baseline performs on par with Avg. Pooling baseline reported in Table 2 and SA cannot obtain a meaningful gain over its counter global average pooling baseline. We also tried (not reported here) solely using LFWA training instances, without pre-training on CelebA, and observed that SA was indeed effective. However it was not able to reach the performance of the model initialized with CelebA. Detailed per-attribute results of our top models for both CelebA and LFWA datasets are shown in Table 5.

4.3 Evaluation of Person Attribute Prediction

Table 3 compares our proposed method with the state-of-the-art on WIDER Attribute [23] dataset. We observe that the Inception-V3 [6] baseline, despite being considerably shallower, performs on par with ResNet-101. Symbiotic Augmentation (SA) which employs semantic segmentation yields a $\sim$2% performance gain over our Inception-V3 [6] baseline surpassing [70], the current state-of-the-art. For detailed performance comparison between varieties of ResNet [71] and DenseNet [72] architectures on WIDER Attribute [23] dataset, readers are encouraged to refer to [70].

Table 4 compares our proposed method with the state-of-the-art on Berkeley Attributes of People [15] dataset. Note that [23] leverages the context in the image while our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

method solely operates on the bounding box of each person, yet it still outperforms [23] with 2.6% margin. Similar to WIDER Attribute [23] dataset, here utilizing semantic segmentation through our proposed Symbiotic Augmentation (SA) results in 2% gain in AP over our already very competitive Inception-V3 [6] baseline. Detailed per-attribute results of our models are shown in Table 6.

4.4 Visualizations

Unlike the global average pooling which equally affects a rather large spatial domain, we expect SSP to generate activations that are semantically aligned. To evaluate our hypothesis, in Figure 6, we show the activations for the top fifty channels of the last convolution layer. Top row corresponds to our basic network with global average pooling, while the bottom row is generated when we replace global average pooling with SSP. We observe that, activations generated by SSP are clearly more localized than those obtained from the global average pooling.

To better understand how attribute prediction and semantic segmentation models have learned their corresponding tasks, we visualize the embedding convolution layers Φ_S and Φ_A (ref. Figure 4) for simultaneously training of CelebA [12] (original image set) with Helen face [43], and WIDER Attribute [23] with LIP [3]. Figure 9 shows how for each facial attribute (vertical axis), network has learned to employ different semantic regions of face (horizontal axis) in order to predict attributes. Note that these weights are learned through back-propagation and are not hard coded, yet they reveal very interesting observations. First, almost all the attributes give "background" the lowest importance, except attribute "Wearing Necklace" which makes sense as neck falls outside the face region and counted as background in Helen face dataset [43]. Second, the learned importance for the majority of attributes are aligned with human expectations. For instance, all the hair-related attributes are inferred with the most attention of the model being paid to the "Hair" region. The same is true for "Big Nose", "Pointy Nose" and "Eyeglasses" as the model learns to focus on the "Nose" region. Figure 7 illustrates Φ_A for the reverse problem where attributes are supposed to improve semantic face parsing. Figure 8a and 8b show the learned weights of the embedding convolution layer for person attribute prediction and human semantic parsing tasks.

We observe that simultaneously training for attribute prediction and semantic segmentation within Symbiotic Augmentation framework, in addition to the performance gains, provides us with meaningful tools to study how a complex deep neural network infers and relate different semantic labels across multiple tasks.

4.5 Attribute Prediction for Semantic Segmentation

In this work, we have established how semantic segmentation can be used to improve person-related attribute prediction. What if we reverse the roles. Can attributes improve semantic parsing problem? To evaluate this, we focus on facial attributes and compare the performance of semantic face parsing on Helen face [43]. We consider three scenarios. First, initializing Inception-V3 [6] backbone with ImageNet [73] pre-trained weights. Second, training a baseline attribute prediction network on CelebA [12] and using the corresponding weights, once training finished, to initialize semantic face parsing network. Third, training facial attribute and semantic face parsing simultaneously through Symbiotic Augmentation (SA) framework. For the sake of simplicity, solely in this experiment, SA only uses the final activations of the CNN backbone instead of concatenating them with intermediate feature maps as shown in Figure 5. We observed that upgrading to full SA model boosts mean class accuracy by ∼5% and also achieves similar mean IoU. Table 7 shows that pre-training on image-level facial attribute annotations delivers a large performance gain over ImageNet based initialization. This shows that there exists an *interrelatedness* between attribute prediction and semantic segmentation. Furthermore, it suggests that while collecting annotations for semantic parsing is laborious and expensive, instead one can use relevant image-level attribute annotations to initialize a semantic parsing model. The last row in each block of the Table 7 demonstrates how training facial attributes and semantic face parsing jointly, through our proposed Symbiotic Augmentation (SA), can further push the performance boundary with significant margin. Therefore, it is easy to see that when few training instances are available, indeed image-level facial attribute labels can serve as an effective source of weak supervision to improve semantic face parsing task. In fact such *interrelatedness* plays a major role in allowing us to successfully unify semantic segmentation and attribute predictions networks (ref. Section 3) without sacrificing the performance. Jointly training on LIP [3] and WIDER Attribute [23], we did not observe meaningful gain in semantic segmentation task on LIP [3]. We hypothesize that, this is due to the fact that LIP [3] itself already has huge (∼30,000 instances) number of training annotations. In order to confirm this, conducting an experiment where only a small portion of LIP [3] training instances are utilized is needed.

5 Conclusion

Aligned with the trend of part-based attribute prediction methods, we proposed employing semantic segmentation to improve person-related attribute prediction. Specifically, we jointly learn attribute prediction and semantic segmentation in order to mainly transfer localization cues from the latter task to the former. To guide the attention of our attribute prediction model to the regions which different attributes naturally show up, we introduced SSP and SSG. While SSP is used to restrict the aggregation procedure of final activations to regions that are semantically consistent, SSG carries the same notion but applies it to the earlier layers. We then demonstrate that there exist a single unified architecture that can mimic the behavior of SSP and SSG, depending on where in the network architecture it is being used. We evaluated our proposed methods on CelebA, LFWA, WIDER Attribute and Berkeley Attributes of People datasets and achieved state-of-the-art performance. We also showed that attributes can improve semantic segmentation (in case of few training instances) when properly used through our Symbiotic Augmentation (SA) framework. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

Method	SSP+SSG	SSP+SSG⋆	Inception -V3: baseline	Symbiotic Aug. (SA)	SSP+SSG	SSP+SSG	SSP+SSG⋆	Inception -V3: baseline	Symbiotic Aug. (SA)	SSP+SSG
Dataset	CelebA	CelebA	CelebA	CelebA	LFWA	CelebA	CelebA	CelebA	CelebA	LFWA
	Classification Accuracy(%)					**Average Precision(%)**				
5 o Clock Shadow	94.50	95.07	94.34	94.62	79.72	80.36	83.96	80.42	81.63	83.61
Arched Eyebrows	83.06	84.56	83.88	84.12	83.74	77.98	81.17	78.93	79.64	73.07
Attractive	82.25	83.28	82.21	82.27	80.89	91.14	92.50	91.18	91.36	83.83
Bags Under Eyes	85.42	86.15	85.26	85.60	85.09	67.68	70.05	67.24	67.96	95.19
Bald	98.79	99.02	98.92	98.95	92.76	76.43	84.03	79.11	79.40	71.09
Bangs	95.51	96.23	95.72	95.86	91.82	93.86	95.54	94.16	94.65	82.46
Big Lips	71.67	72.45	71.35	72.16	80.20	62.85	62.97	62.30	63.01	81.83
Big Nose	84.50	85.38	84.77	85.01	84.67	68.62	72.25	69.13	71.43	95.92
Black Hair	90.06	90.63	89.96	90.15	92.81	89.75	90.79	89.55	90.13	77.13
Blond Hair	95.82	96.30	95.90	95.94	97.72	91.45	92.73	91.54	91.67	78.77
Blurry	95.67	96.44	95.65	95.85	87.49	53.61	65.87	53.95	57.03	63.88
Brown Hair	89.25	89.95	88.42	88.46	82.72	76.58	78.97	75.22	75.18	83.76
Bushy Eyebrows	92.36	93.20	92.34	92.50	85.77	76.47	81.00	76.36	76.91	94.45
Chubby	95.61	96.02	95.80	95.94	77.66	56.24	62.54	59.63	62.39	76.48
Double Chin	96.28	96.61	96.23	96.47	81.86	58.42	63.92	58.49	61.86	85.80
Eyeglasses	99.27	99.67	99.51	99.48	92.79	98.43	99.20	98.52	98.49	86.96
Goatee	97.28	97.58	97.41	97.55	84.08	74.89	81.64	79.08	80.86	75.74
Gray Hair	98.22	98.37	98.16	98.30	89.24	77.32	80.49	77.65	79.32	71.69
Heavy Makeup	90.83	92.17	91.03	90.99	95.90	96.26	97.31	96.29	96.30	88.80
High Cheekbones	87.13	88.13	87.09	87.48	89.48	94.94	95.78	94.92	95.23	91.68
Male	97.67	98.51	98.00	98.08	94.42	99.59	99.83	99.69	99.73	99.08
Mouth Slightly Open	92.25	94.19	92.61	92.79	84.29	97.97	98.87	98.10	98.29	88.36
Mustache	96.96	97.01	96.94	97.16	94.01	64.14	67.94	65.45	67.01	86.11
Narrow Eyes	86.68	87.92	86.86	87.17	84.68	52.35	59.31	53.22	55.11	95.22
No Beard	95.66	96.52	95.77	95.74	83.63	99.74	99.82	99.76	99.79	94.98
Oval Face	77.83	76.83	77.15	77.50	77.89	66.25	63.84	65.40	65.75	87.21
Pale Skin	97.08	97.29	96.78	96.69	91.15	67.25	70.65	60.60	60.32	97.77
Pointy Nose	76.50	77.86	77.14	77.45	84.99	60.67	65.93	62.74	63.67	95.69
Receding Hairline	93.31	94.14	93.42	93.81	86.60	60.24	67.80	62.05	63.79	95.57
Rosy Cheeks	94.78	95.39	94.75	94.77	86.28	67.66	72.40	64.33	65.41	74.02
Sideburns	97.70	98.00	97.75	97.82	83.21	82.92	86.78	83.16	85.17	81.54
Smiling	91.92	93.39	92.00	92.45	92.51	97.97	98.62	98.07	98.23	97.00
Straight Hair	83.59	84.46	85.16	85.21	81.58	63.56	66.22	68.82	69.21	83.26
Wavy Hair	84.79	84.62	86.13	85.93	81.22	88.46	88.73	90.15	90.27	87.69
Wearing Earrings	89.99	90.94	90.41	90.56	95.23	83.40	85.71	84.79	85.18	89.11
Wearing Hat	98.78	99.11	99.07	99.07	91.08	92.87	95.89	95.21	95.59	75.11
Wearing Lipstick	93.58	94.56	93.61	93.88	95.19	98.67	99.10	98.70	98.76	90.52
Wearing Necklace	88.72	88.01	89.65	89.57	90.15	59.05	52.89	62.92	62.71	82.38
Wearing Necktie	97.15	97.02	97.17	97.12	83.87	86.81	87.51	87.45	88.31	94.47
Young	87.85	89.01	88.52	88.37	86.95	96.89	97.60	97.13	97.19	74.02
Avg.	91.16	91.80	91.32	91.47	87.13	78.74	81.45	79.28	80.10	85.28

TABLE 5: Detailed per-attribute classification accuracy(%) and average precision(%) results of our proposed models for facial attribute prediction. Note that SSP+SSG⋆ indicates the experiment using pre-cropped images of CelebA.

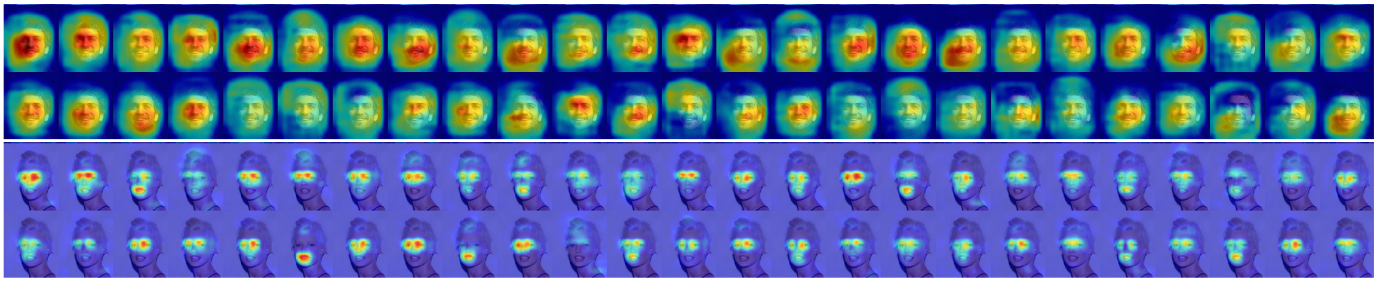


Fig. 6: Top fifty activation maps of the last convolution layer sorted in descending order w.r.t the average activation values. Top: Basic attribute prediction model using global pooling. Bottom: SSP.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

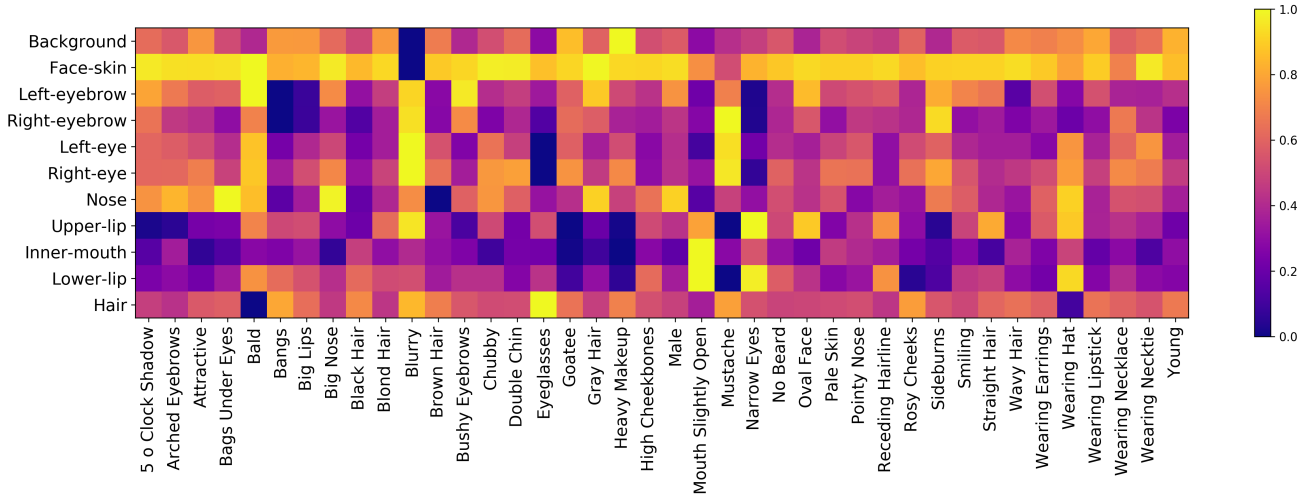


Fig. 7: Learned weights of Φ_A in Symbiotic Augmentation (SA), trained on CelebA and Helen. Note: 9 values associated with 3×3 kernels are averaged. For better visualization, values in each row are normalized between 0 and 1.

WIDER Attribute [23]		
	Inception-V3: baseline	Symbiotic Augmentation (SA)
Male	95.60	96.64
Long Hair	86.98	89.25
Sunglasses	70.56	78.31
Hat	92.87	95.04
T-shirt	83.36	84.77
Long Sleeve	96.71	97.64
Formal	83.82	85.38
Shorts	91.96	93.87
Jeans	79.60	81.76
Long Pants	97.18	97.74
Skirt	85.74	87.65
Face Mask	76.51	79.18
Logo	91.07	90.87
Stripe	70.15	68.04
Avg.	85.86	87.58
Berkeley Attributes of People [15]		
	Inception-V3: baseline	Symbiotic Augmentation (SA)
Is Male	96.29	96.73
Has Long Hair	93.71	94.41
Has Glasses	79.57	88.41
Has Hat	92.97	96.31
Has T-shirt	86.28	88.15
Has Long sleeves	96.96	98.01
Has Shorts	95.43	95.82
Has Jeans	95.34	95.80
Has Long Pants	99.33	99.55
Avg.	92.87	94.80

TABLE 6: Detailed per-attribute AP(%) results of our proposed models for person attribute prediction.

hope to encourage future research works to invest more in the interrelatedness of these two problems.

REFERENCES

[1] M. M. Kalayeh, B. Gong, and M. Shah, "Improving facial attribute prediction using semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6942–6950. 1, 2, 3, 4, 5, 7, 10

[2] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3484–3491. 2, 5, 8, 9, 14

[3] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 932–940. 2, 5, 8, 9, 11

[4] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. 2

[5] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528. 2, 5

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. 3, 5, 7, 8, 10, 11

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1778–1785. 3

[8] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 951–958. 3

[9] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2352–2359. 3

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 365–372. 4

[11] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European conference on computer vision*. Springer, 2008, pp. 340–353. 4, 9, 10

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 4, 8, 9, 10, 11

[13] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3337–3344. 4

[14] T. Berg and P. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 955–962. 4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence

Intersection over Union(%)												
Method	bkg	face skin	l-eyebrow	r-eyebrow	l-eye	r-eye	nose	u-lip	i-mouth	l-lip	hair	**Avg.**
init: ImageNet	92.97	85.58	46.90	48.33	55.39	55.91	84.24	43.77	59.21	55.19	71.99	63.58
init: CelebA	93.20	86.40	51.31	51.11	56.22	58.81	84.82	49.32	60.01	58.95	73.13	65.75
SA	**94.25**	**88.24**	**59.29**	**58.11**	**62.45**	**67.22**	**87.96**	**51.05**	**69.66**	**70.32**	**75.77**	**71.29**
Class Accuracy(%)												
method	bkg	face skin	l-eyebrow	r-eyebrow	l-eye	r-eye	nose	u-lip	i-mouth	l-lip	hair	**Avg.**
init: ImageNet	96.04	94.21	56.02	60.95	67.61	67.62	90.69	58.25	74.73	66.12	83.36	74.14
init: CelebA	95.96	94.09	63.31	67.71	67.30	69.79	90.06	66.80	75.27	72.83	85.22	77.12
SA	**97.02**	**95.47**	**69.89**	**74.97**	**72.12**	**77.21**	**92.43**	**66.96**	**76.88**	**81.60**	**84.67**	**81.07**

TABLE 7: Effect of leveraging image-level attribute supervision for semantic face parsing, evaluated on the test split of Helen face [43] [2]. Here, all the models were trained with the input image resolution of 448×448.

[15] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1543–1550. 4, 8, 10, 13

[16] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *European conference on computer vision*. Springer, 2012, pp. 609–623. 4

[17] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1761–1768. 4

[18] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1629–1636. 4

[19] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss *et al.*, "Understanding objects in detail with fine-grained attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3622–3629. 4

[20] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *European Conference on Computer Vision*. Springer, 2010, pp. 155–168. 4

[21] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1681–1688. 4

[22] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 87–97. 4

[23] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700. 4, 8, 9, 10, 11, 13

[24] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2470–2478. 4, 10

[25] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644. 4, 9, 10

[26] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 753–769. 4

[27] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025. 4

[28] E. Rudd, M. Günther, and T. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," *arXiv preprint arXiv:1603.07027*, 2016. 4, 8, 9, 10

[29] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5375–5384. 4, 8, 9, 10

[30] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1851–1860. 4, 8, 9, 10

[31] J. Li, F. Zhao, J. Feng, S. Roy, S. Yan, and T. Sim, "Landmark free face attribute prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4651–4662, 2018. 4, 9, 10

[32] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8. 4

[33] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013. 4

[34] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851. 4

[35] 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, pp. 3431–3440. 4, 5

[36] 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*, vol. 39, no. 12, pp. 2481–2495, 2017. 4, 5

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. 4, 5

[38] 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*, vol. 40, no. 4, pp. 834–848, 2018. 4, 5

[39] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations*, 2015. 4, 5

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 5

[41] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation," *arXiv preprint arXiv:1611.06612*, 2016. 5

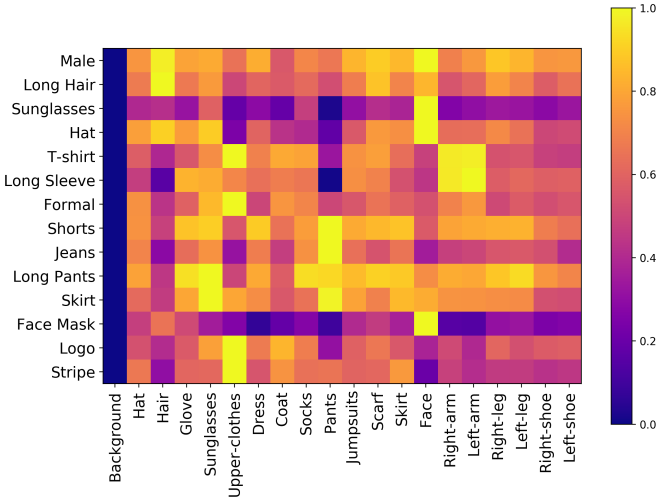
[42] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 5

[43] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692. 5, 8, 9, 11, 14

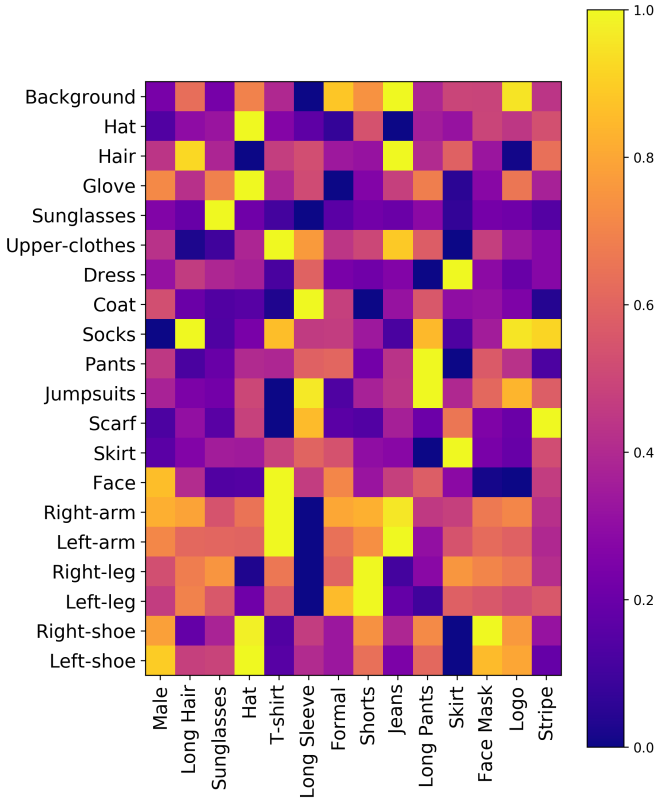
[44] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting crfs with boltzmann machine shape priors for image labeling," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2019–2026. 5

[45] S. Liu, J. Yang, C. Huang, and M.-H. Yang, "Multi-objective convolutional learning for face labeling," in *Proceedings of the IEEE*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2956039, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Φ_S



(b) Φ_A

Fig. 8: Learned weights of embedding convolution layers in Symbiotic Augmentation (SA), trained on WIDER and LIP. Note: 9 values associated with 3×3 kernels are averaged. For better visualization, values in each row are normalized between 0 and 1.

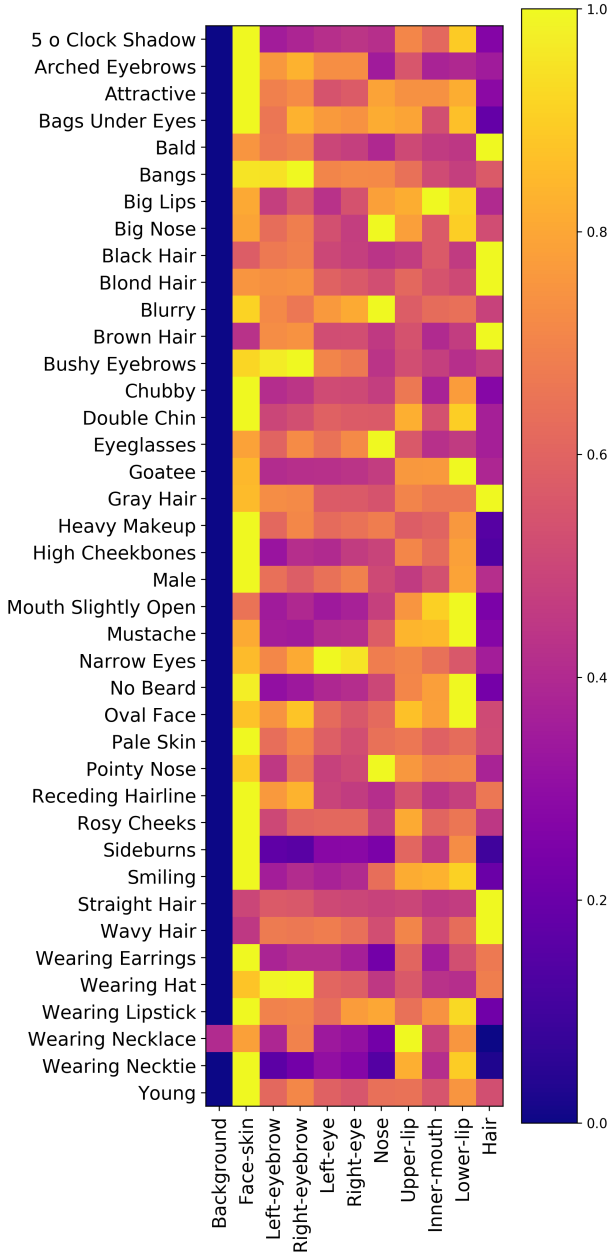


Fig. 9: Learned weights of Φ_S in Symbiotic Augmentation (SA), trained on CelebA and Helen. Note: 9 values associated with 3×3 kernels are averaged. For better visualization, values in each row are normalized between 0 and 1.

Conference on Computer Vision and Pattern Recognition, 2015, pp. 3451–3459. 5

[46] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[47] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Joint object and part segmentation using deep learned potentials," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1573–1581. 5

[48] B. Hariharan, P. Arbeláez, 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*, 2015, pp. 447–456. 5

[49] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *European Conference on Computer*

Vision. Springer, 2016, pp. 125–143. 5

[50] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649. 5

[51] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, "Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net," in *European Conference on Computer Vision*. Springer, 2016, pp. 648–663. 5

[52] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3570–3577. 5

[53] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2402–2414, 2015. 5

[54] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1386–1394. 5

[55] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3408–3415. 5

[56] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, and S. Yan, "Fashion parsing with video context," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1347–1358, 2015. 5

[57] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3182–3189. 5

[58] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3519–3526. 5

[59] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-cnn meets knn: Quasi-parametric human parsing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1419–1427. 5

[60] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016. 6

[61] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015. 6, 9

[62] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1600–1609. 8

[63] J. Wang, Y. Cheng, and R. S. Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," *arXiv preprint arXiv:1604.06433*, 2016. 9, 10

[64] Y. Zhong, J. Sullivan, and H. Li, "Leveraging mid-level deep representations for predicting face attributes in the wild," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3239–3243. 9, 10

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361. 9

[66] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. 10

[67] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088. 10

[68] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint arXiv:1707.06089*, 2017. 10

[69] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5513–5522. 10

[70] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *The European Conference on Computer Vision (ECCV)*, September 2018. 10

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 10

[72] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3. 10

[73] 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," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. 11