# Improving Facial Attribute Prediction using Semantic Segmentation

Mahdi M. Kalayeh
Mahdi@eecs.ucf.edu

Boqing Gong
bgong@crcv.ucf.edu

Mubarak Shah
shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida

## Abstract

*Attributes are semantically meaningful characteristics whose applicability widely crosses category boundaries. They are particularly important in describing and recognizing concepts where 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 facial attribute prediction. The core idea lies in the fact that many facial attributes describe local properties. In other words, the probability of an attribute to appear in a face image is far from being uniform in the spatial domain. We build our facial 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 recognition, we are able to localize the attributes, despite merely having access to image level labels (weak supervision) during training. We evaluate our proposed method on CelebA and LFWA datasets and achieve superior results to the prior arts. Furthermore, we show that in the reverse problem, semantic face parsing improves when facial attributes are available. That reaffirms the need to jointly model these two interconnected tasks.*

## 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 examples 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 the *sharable concepts*. They are human describable and machine detectable. The fact that attributes are generally not category-specific suggests that one can potentially describe an exponential number of categories with various combinations of attributes. Naturally, attributes are "additive" to the objects (*e.g.*, horn for cow). 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 cow with or without horn is still a cow). 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.*, the horn, if appears, would show up on a cow's head). This notion is the basis of our work. That is, in order to detect an attribute, instead of the entire spatial domain, we should focus on the region in which that attribute naturally shows up. We hypothesize that the attribute prediction can benefit from localization cues. However, attribute prediction benchmarks come with holistic image level labels. In addition, sometimes it is hard to define a spatial boundary for a given attribute. For instance, it is not clear that according to which spatial region in a face one decides if a person is "attractive" or not. To tackle this challenge, we transfer localization cues from a relevant auxiliary task to the attribute prediction problem.

Using bounding box to show the boundary limits of an object is a common practice in computer vision. However, regions that different attributes occupy drastically change in shape and form. 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. Therefore, we need an auxiliary task that learns detailed localization information without restricting the corresponding regions to be in 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, we are interested in facial attributes. Hence, the semantic face parsing problem [21] is a suitable candidate to serve as an auxiliary task to spatially hint the attribute prediction methods.

To perform attribute prediction, we feed an image to a fully convolutional neural network which generates feature maps that are ready to be aggregated [15] and passed to the classifier. However, global pooling [15] 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 them 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 segmentation into earlier layers of the attribute prediction network with a gating mechanism. Specifically, we augment the max pooling operation such that it does not mix activations that reside in different semantic regions. To do so, we gate the activation output of the last convolution layer prior to the max pooling by element-wise multiplying it with the semantic regions. This 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 segmentation is not available for the attribute benchmarks, we learn to *estimate* it using a deep semantic segmentation network. Our approach is conceptually similar to [17] in which an encoder-decoder model is built using convolution and deconvolution layers. However, considering the relatively small number of available data for the auxiliary segmentation problem, we modify the network architecture in order to adapt it to our facial attribute prediction problem. Despite being much simpler than [17], we found our semantic segmentation network to be very effective in solving the auxiliary task of semantic face parsing. Once trained, such network is able to provide localization cues in the form of semantic segmentation (decoder output) that decompose the spatial domain of an image into mutually exclusive semantic regions.

We show that both SSP and SSG mechanisms outperform the existing state-of-the-art facial attribute prediction techniques while employing them together results in further improvements.

## 2. Related Work

It is fair to say that the attribute prediction literature can be divided into holistic and part-based approaches. The common theme among the holistic methods is to take the entire image into account when extracting features for attribute prediction. On the other hand, part-based methods begin with an attribute-related part detection and then use the localized parts, in isolation from the rest of the image, to extract features.

Our proposed method falls between the two ends of the spectrum. While we process the image in a holistic fashion to generate feature vectors for the classifiers, we employ localization cues in the form of semantic segmentation.

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 extracted features. Among earlier works we refer to [13, 1, 3] as successful examples of part-based attribute prediction approaches. More recently, in an effort to combine part-based models with deep learning, Zhang *et al.* [23] proposed PANDA, a pose-normalized convolutional neural network (CNN) to infer human attributes from images. PANDA employs poselets [3] to localize body parts and then extracts CNN features from the localized regions. These features will later be used to train SVM classifiers for attribute prediction. Inspired by [23] while seeking to also leverage the holistic cues, Gkioxari *et al.* [5] proposed a unified framework that benefits from both holistic and part-based clues while utilizing a deep version of poselets [3] as part detectors. Liu *et al.* [16] 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 located region where they perform facial attribute prediction. Finally, in a part-based approach, Singh *et al.* [20] use spatial transformer networks [10] to locate 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.

## 3. Methodology

In this section, we begin with the attribute prediction models assuming that the semantic regions are given. We then move on to the semantic segmentation network and provide details on how the semantic regions are generated.

### 3.1. Attribute Prediction Networks

To leverage the localization cues for facial attribute prediction, we propose semantic segmentation-based pooling and gating mechanisms. We describe our basic attribute prediction model. Then, we explain SSP and SSG in detail including how they are employed in the basic model, simply as new layers, to improve facial attribute prediction.

#### 3.1.1 Basic Attribute Prediction Network

Our basic attribute prediction model is a 12-layers deep fully convolutional neural network. We gradually increase the number of convolution filters from 64 to 1024 filters as we proceed towards the deeper layers. Prior to any increase in the number of convolution filters, we reduce the size of the activation maps using max pooling. For such operation both the kernel size and stride values are set to 2. In our architecture, every convolution layer is followed by the Batch Normalization [9] and PReLU [7]. The kernel size and stride values of all the convolution layers are respectively set to 3 and 1. The first 8 layers of our basic attribute prediction network are similar in configuration to the encoder part of the semantic segmentation network and detailed in Table 1. The rest consists of 4 convolution layers of 512 and 1024 filters, two layers of each. At the end of the pipeline, we aggregate the activations of the last convolution layer using global average pooling [15] to generate 1024-D vector representations. These vectors are subsequently passed to the classifier for attribute prediction. We train the network using sigmoid cross entropy loss. Section 5 provides further details on the training procedure.

#### 3.1.2 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 [15] used above in our basic attribute prediction model 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 1024-D vector representation, we obtain multiple features, each representing only a single semantic region. This approach has an interesting intuition behind it. In fact, SSP funnels the backpropagation 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.

While we can simply concatenate the representations associated with different regions and pass it to the classifier, 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 approach to [2] where Bilen and Vedaldi employed a two branch network for weakly supervised object detection.

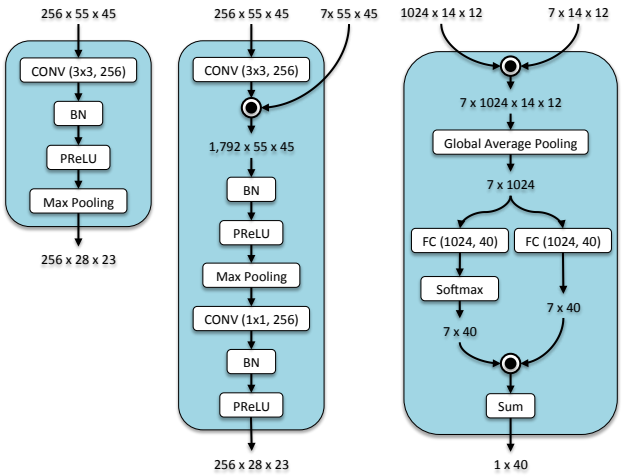


Figure 1. Left: Standard convolution layer followed by max pooling, Middle: SSG, Right: SSP. Note: In this work, there are 7 semantic regions and 40 attributes to predict.

We pass the vector representations, each associated to a different semantic region, to two branches one for recognition and another for localization. We implement these branches as linear classifiers that map 1024-D vectors 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 begin by normalizing the output 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 score by a weighted sum of the recognition branch outputs using weights generated by the localization branch. Figure 1, on the right, shows the SSP architecture.

#### 3.1.3 SSG: Semantic Segmentation-based Gating

The 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. Unlike that, many attributes are inherently localized to 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 1 shows a standard convolution layer followed by max pooling on the left, and the SSG architecture in the middle. The latter is our proposed alternative to the former. Here we assume the convolution layer to preserve the number of input channels but it does not have to be. To gate

the output activations of the convolution layer, we broadcast element-wise multiplication for each of the $N = 7$ semantic regions with the entire activation maps. This generates $N$ copies (totally $1,792 = 256 \times 7$ activation maps) of the activations that are masked differently. Such mechanism spatially decomposes the activation maps into copies where activations with high values cannot simultaneously occur in two semantically different regions. For example, gating with the semantic segmentation 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 sum up to 1. We then gate the activations right after the convolution and before the Batch Normalization [9]. This is very important since the Batch Normalization [9] 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 $1\times1$ convolution. It is worth noting that SSG can mimic the standard max pooling by learning a sparse set of weights for the $1\times1$ convolution. In a nutshell, semantic segmentation-based gating allows us to process the activations of convolution layers in a per-semantic region fashion, and directly learns how to combine the pooled values afterwards.

### 3.2. Semantic Segmentation Network

We have previously explained the rationale behind employing semantic face parsing to improve facial attribute prediction. Our design for the semantic segmentation network follows an encoder-decoder approach, similar in concept to the deconvolution network proposed in [17]. However, considering the limited number of training data for the segmentation network, we have made different design decisions to reduce the complexity of the model while preserving its capabilities. The encoder consists of 8 convolution layers in blocks of 2, separated with 3 max pooling layers. This is much smaller than the 13 layers used in the deconvolution network [17]. At the end of the encoder part, rather than collapsing the spatial resolution as in [17], we maintain it at the scale of one-eighth of the input size. The decoder is a mirrored version of the encoder replacing convolution layers with deconvolution and max pooling layers with upsampling. Unlike [17] that uses switch variables to store the max pooling locations, we simply upsample the activation maps (repetition with nearest neighbor interpolation). We increase (decrease) the number of convolution (deconvolution) filters by a factor of 2 after each max pooling (upsampling), starting from 64 (512) filters as we proceed along the encoder (decoder) path. Every convolution and deconvolution layer is followed by Batch Normalization [9] and PReLU [7]. To cope with the challenge of relatively small number of training data, we propagate the semantic segmentation loss at different depths along the decoder path. That is, before each upsampling layer, we compute the loss by predicting the semantic segmentation maps at different scales. We then aggregate these losses with equal weights prior to backpropagation. Finally, while [17] employs VGG16 [19] weights to initialize the encoder, we train our network from scratch. These design decisions allow us to successfully train the semantic segmentation network with the limited number of training data. Detailed configuration of the semantic segmentation network is shown in Table 1.

| Layer | Operations | Output size |
|---|---|---|
| $\text{Conv}_{11}$ | Conv, BN, PReLU | $64\times218\times178$ |
| $\text{Conv}_{12}$ | Conv, BN, PReLU | $64\times218\times178$ |
| $\text{MaxPool}_{1}$ | Max Pooling | $64\times109\times89$ |
| $\text{Conv}_{21}$ | Conv, BN, PReLU | $128\times109\times89$ |
| $\text{Conv}_{22}$ | Conv, BN, PReLU | $128\times109\times89$ |
| $\text{MaxPool}_{2}$ | Max Pooling | $128\times55\times45$ |
| $\text{Conv}_{31}$ | Conv, BN, PReLU | $256\times55\times45$ |
| $\text{Conv}_{32}$ | Conv, BN, PReLU | $256\times55\times45$ |
| $\text{MaxPool}_{3}$ | Max Pooling | $256\times28\times23$ |
| $\text{Conv}_{41}$ | Conv, BN, PReLU | $512\times28\times23$ |
| $\text{Conv}_{42}$ | Conv, BN, PReLU | $512\times28\times23$ |
| $\text{Deconv}_{41}$ | Deconv, BN, PReLU | $512\times28\times23$ |
| $\text{Deconv}_{42}$ | Deconv, BN, PReLU | $512\times28\times23$ |
| $\text{UpSample}_{3}$ | UpSampling | $512\times55\times45$ |
| $\text{Deconv}_{31}$ | Deconv, BN, PReLU | $256\times55\times45$ |
| $\text{Deconv}_{32}$ | Deconv, BN, PReLU | $256\times55\times45$ |
| $\text{UpSample}_{2}$ | UpSampling | $256\times109\times89$ |
| $\text{Deconv}_{21}$ | Deconv, BN, PReLU | $128\times109\times89$ |
| $\text{Deconv}_{22}$ | Deconv, BN, PReLU | $128\times109\times89$ |
| $\text{UpSample}_{1}$ | UpSampling | $128\times218\times178$ |
| $\text{Deconv}_{11}$ | Deconv, BN, PReLU | $64\times218\times178$ |
| $\text{Deconv}_{12}$ | Deconv, BN, PReLU | $64\times218\times178$ |
| $\text{Deconv}_{13}$ | Deconv, BN, PReLU | $7\times218\times178$ |

Table 1. Configuration of the Semantic Segmentation Network. For all the convolution/ deconvolution layers, kernel size and stride values are respectively set to 3 and 1. To prevent confusion, we are not showing the side loss layers, namely $\text{Deconv}_{43}$, $\text{Deconv}_{33}$ and $\text{Deconv}_{23}$.

## 4. Experimental Results

### 4.1. Training Semantic Segmentation Network

In this paper, we are interested in facial attribute prediction. Hence, face parsing problem [21] which aims at pixel-level classification of a face image into multiple se-

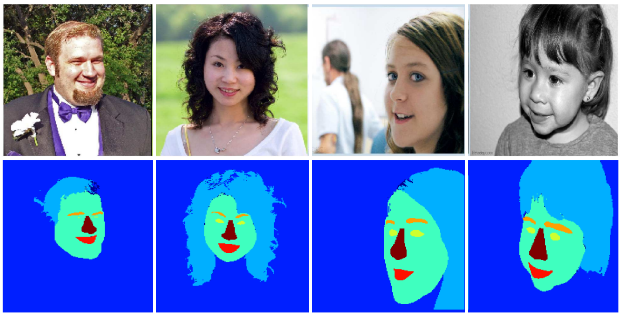

Figure 2. Examples of the Helen face dataset [14] supplemented with segment label annotations [21] and then grouped into 7 semantic classes. In bottom row, colors indicate different class labels.

mantic regions is a suitable auxiliary task for us. To train the semantic segmentation network, we begin with 11 segment label annotations per image that [21] provides to supplement Helen face dataset [14]. These labels are as follows: background, face skin (excluding ears and neck), left eyebrow, right eyebrow, left eye, right eye, nose, upper lip, inner mouth, lower lip and hair. We combine left and right eye (eyebrow) labels to create a single eye (eyebrow) label. Similarly, we aggregate upper lip, inner mouth, and lower lip to generate a single mouth label. As a result we end up with a total of 7 labels (background, hair, face skin, eyes, eyebrows, mouth and nose). Figure 2 illustrates a few instances of the input images along with their corresponding segment label annotations. The face parsing dataset [21] comes with 2,330 images in three splits of 2000, 230 and 100, respectively for training, validation and test. However, for the attribute prediction task, we can use the entire dataset to train the semantic segmentation network. We train our model with softmax cross entropy loss. Section 5 provides details on the training procedure. Figure 3 shows a few examples of segmentation maps generated by our network. Despite very few number of training data used in its training process, the semantic segmentation network is able to successfully localize various facial regions in previously unseen images. Later, we evaluate our proposed attribute prediction model where these semantic segmentation cues are utilized to improve facial attribute prediction.

### 4.2. Datasets and Evaluation Metrics

We mainly evaluate our proposed approach on the CelebA dataset [16]. CelebA 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. Images are annotated with 40 facial attributes such as, "wavy hair", "mouth slightly open", "big lips", etc. In addition to the original images, CelebA provides a set of pre-cropped images. We report our results on both of these

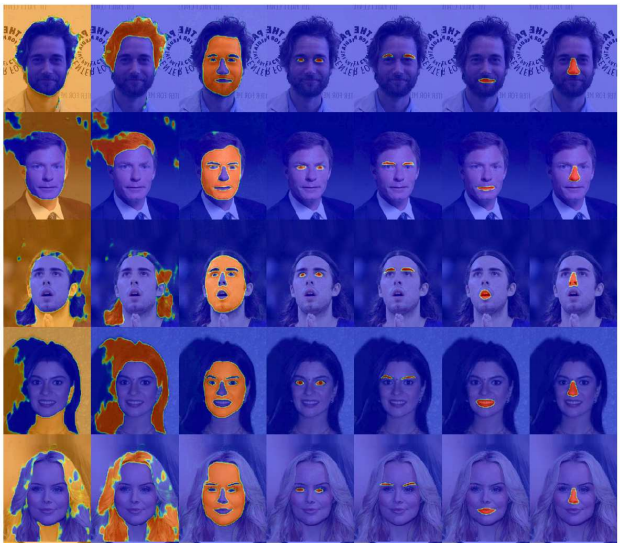

Figure 3. Examples of the segmentation masks generated by our semantic segmentation network for previously unseen images. From left to right: background, hair, face skin, eyes, eyebrows, mouth and nose.

image sets. It is worth noting that Liu *et al.* [16] have used both the training and validation data in order to train different parts of their model. In particular, training data has been used to pre-train and fine-tune ANet and LNet while they train SVM classifiers using the validation data. In our experiments, we only use the training split to train our attribute prediction networks.

To supplement the analyses on CelebA dataset [16], we also provide experimental results on LFWA[16]. LFWA has a total of 13,232 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[16] dataset. For the LFWA dataset [16], we follow the same evaluation protocol as the one for CelebA dataset [16].

To evaluate the attribute prediction performance, Liu *et al.* [16] use classification accuracy/error. 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 [18, 8] as well. Therefore, in addition to the classification error, we also report the average precision of the prediction scores.

### 4.3. Evaluation of Facial Attribute Prediction

For all the numbers reported here, we want to point out that FaceTracer [12] and PANDA [23] use groundtruth landmark points to attain face parts. Wang *et al.* [22] use 5 million auxiliary image pairs, collected by the authors, to pretrain their model. Wang *et al.* [22] 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 mod-

els from scratch with only attribute labels and the auxiliary face parsing labels.

#### 4.3.1 Evaluation on CelebA dataset

We compare our proposed method with the existing state-of-the-art attribute prediction techniques on the CelebA dataset [16]. To prevent any confusion and have a fair comparison, Table 2 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. We see that even our basic model with global average pooling, with the exception of the MOON [18], outperforms previous state-of-the-art techniques. Accordingly, we can make two observations.

First, a simple yet well designed architecture can be very effective. Liu *et al.* [16] combine three deep convolutional neural networks with SVM and Rudd *et al.* [18] have adopted VGG16 [19] topped with a novel objective function. These models are drastically larger than our basic network. Specifically, in [16], $\text{LNet}_o$ and $\text{LNet}_s$ have network structures similar to AlexNet [11]. AlexNet has 60M parameters. Thus, only the localization part in [16], not considering ANet, has a total of 120M parameters. Rudd *et al.* [18] adopt VGG16 [19] that has 138M parameters. Our basic attribute prediction network has only 24M parameters thanks to replacing fully connected layers with a single global average pooling.

Second, [18] and [16] are built on the top of networks previously trained on massive object category (and facial identity) data while we train all our networks from scratch. Hence, we reject the necessity of pre-training on other large scale benchmarks, arguing that CelebA dataset [16] itself is sufficiently large for successfully training facial attribute prediction models from scratch.

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 baseline already beats the majority of 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 [6]. 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 2 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 [12] | 18.88 | – |
| PANDA [23] | 15.00 | – |
| Liu *et al.* [16] | 12.70 | – |
| Wang *et al.* [22] | 12.00 | – |
| Zhong *et al.* [24] | 10.20 | – |
| Rudd *et al.* [18]: Separate | – | 9.78 |
| Rudd *et al.* [18]: MOON | – | 9.06 |
| SPPNet* | – | 9.49 |
| Naive Approach | 9.62 | 9.13 |
| BBox | – | 8.76 |
| Ours: Avg. Pooling | 9.83 | 9.14 |
| Ours: SSG | 9.13 | 8.38 |
| Ours: SSP | 8.98 | 8.33 |
| Ours: SSP + SSG | 8.84 | 8.20 |
| **Average Precision%** | | |
| Method | Original | Pre-cropped |
| SPPNet* | – | 77.69 |
| Naive Approach | 76.29 | 79.74 |
| BBox | – | 79.95 |
| Ours: Avg. Pooling | 77.16 | 79.74 |
| Ours: SSG | 77.46 | 80.55 |
| Ours: SSP | 78.01 | 81.02 |
| Ours: SSP + SSG | 78.74 | 81.45 |
| **Balanced Accuracy% [8]** | | |
| Method | Original | Pre-cropped |
| Huang *et al.* [8] | – | 84.00 |
| Ours: Avg. Pooling | – | 86.73 |
| Ours: SSG | – | 87.82 |
| Ours: SSP | – | 88.24 |

Table 2. Attribute prediction performance evaluated by the classification error, average precision and balanced classification accuracy [8] on the CelebA [16] 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 [9]. We train the network using the same setting as other aforementioned mod-

| Method | Classification Error% | AP% |
|---|---|---|
| FaceTracer [12] | 26.00 | – |
| PANDA [23] | 19.00 | – |
| Liu *et al.* [16] | 16.00 | – |
| Zhong *et al.* [24] | 14.10 | – |
| Wang *et al.* [22] | 13.00 | – |
| Ours: Avg. Pooling | 14.73 | 82.69 |
| Ours: SSG | 13.87 | 83.49 |
| Ours: SSP | 13.20 | 84.53 |
| Ours: SSP + SSG | 12.87 | 85.28 |

Table 3. Attribute prediction performance evaluated by the classification error and the average precision (AP) on LFWA [16] dataset.

els. Our experimental results indicate that such naive approach cannot leverage the localization cues as good as our proposed methods. Table 2 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 *addition* fashion. Instead, our proposed mechanisms take a *multiplication* approach by masking the activations using the semantic regions.

**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 regions 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 [16], 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 2, 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 [8, 18], we suggested using average precision instead of classification accuracy/error to evaluate attribute prediction. Instead, Huang *et al.* [8] have adopted balanced accuracy measure. To see if our proposed approach is superior to [8] under balanced accuracy measure, we fine-tuned our models with the weighted ($\propto$ imbalance level) binary cross entropy loss. From Table 2, we observe that under balanced accuracy [8], all the variations of our proposed model outperform [8] with large margins.

| Region | w/o Attributes | w/ Attributes |
|---|---|---|
| Bakcground | 89.25 | 89.64 |
| Hair | 47.56 | 48.32 |
| Face skin | 78.65 | 79.92 |
| Eyes | 46.83 | 56.33 |
| Eyebrows | 31.22 | 42.25 |
| Mouth | 62.03 | 65.42 |
| Nose | 77.40 | 77.74 |
| **Average** | 61.84 | 65.66 |

Table 4. Effect of facial attributes on semantic face parsing performance evaluated by Intersection over Union (IoU%).

#### 4.3.2 Evaluation on LFWA dataset

To better understand the effectiveness of our proposed approach, we report experimental results on the LFWA dataset [16] in Table 3. We observe that, all the models proposed in this work which exploit localization cues improve our basic model. Specifically, SSP + SSG achieves considerably better performance than the average pooling basic model with 1.86% in classification error and 2.59% in the average precision. Our best model also outperforms all other state-of-the-art methods.

### 4.4. Facial Attributes for Semantic Face Parsing

In this work, we established how semantic segmentation can be used to improve facial attribute prediction. What if we reverse the roles. Can facial attributes improve semantic face parsing? To evaluate this, we jointly train two networks where the first 8 layers of our basic attribute prediction network share weights with the encoder part of the semantic segmentation network. We optimize w.r.t the aggregation of two losses. Specifically, the attribute prediction loss on the CelebA [16] dataset and the semantic segmentation loss on the Helen face [14] dataset using facial segment labels of [21]. We follow pre-defined data partitions of [21], detailed in section 4.1, and use Intersection over Union (IoU) as the evaluation measure. Table 4 shows nearly 4% boost when attributes are incorporated, indicating the positive effect of attributes in improving semantic face parsing. This shows that there exist an interrelatedness between attribute prediction and semantic segmentation. In future, we will further explore this promising direction.

### 4.5. Visualizations

Figure 4 illustrates per-attribute weights that the localization branch of the SSP has learned in order to combine the predictions associated with different semantic regions. We observe that attributes such as “Black Hair”, “Brown Hair”, “Straight Hair” and “Wavy Hair” have strong bias towards the hair region. This matches our expectation. However, attribute “Blond Hair” does not behave similarly. We suspect

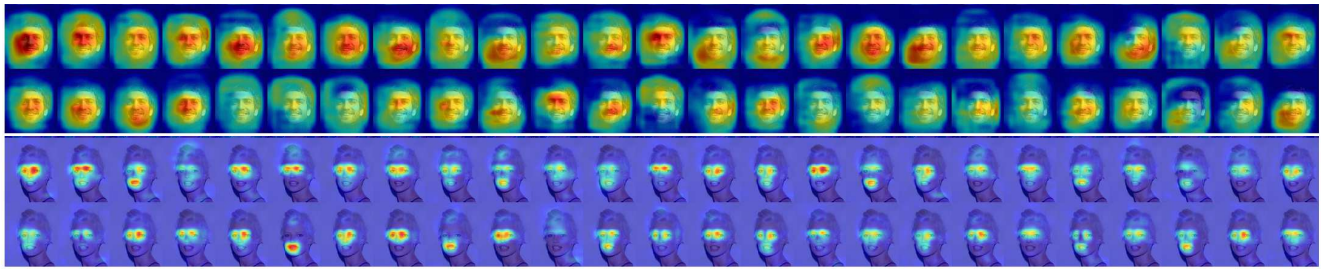

Figure 5. Top fifty activation maps of the last convolution layer sorted in descending order w.r.t the average activation values. Top: Global average pooling. Bottom: SSP.

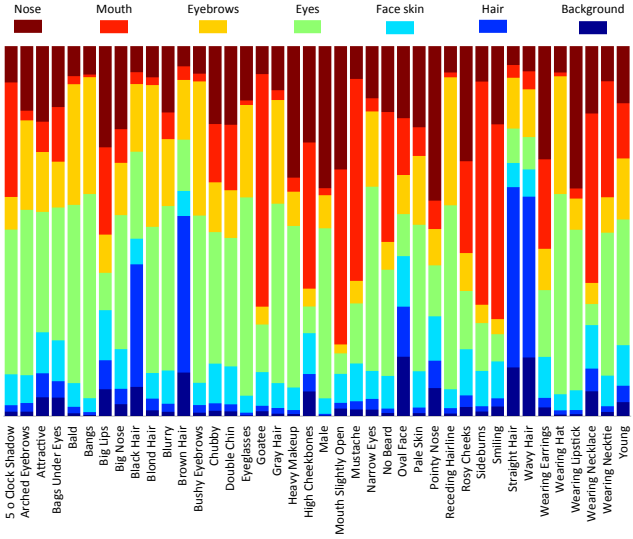


Figure 4. Contribution of different semantic regions in attribute prediction as learned by the localization branch of SSP. Values are averaged over multiple random mini-batches of 32 images.

that it is because the semantic segmentation network does not perform as consistent on light hair colors as it does on the dark ones (refer to Figure 3). Attributes such as "Goatee", "Mouth Slightly Open", "Mustache" and "Smiling" are also showing a large bias towards the mouth region. While these are aligned with our human knowledge, "Sideburns" and "Wearing Necklace" apparently have incorrect biases. Unlike the global 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 5, 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.

## 5. Implementation Details

All of our experiments were conducted on a single NVIDIA Titan X GPU. We use AdaGrad [4] with mini-batches of size 32 to train the attribute prediction models from scratch. The learning rate and weight decay are respectively set to 0.001 and 0.0005. We follow the same setting for training the semantic segmentation network. We perform data augmentation by randomly flipping (horizontally) the input images. In SSP experiments, we resize the output of the semantic segmentation network at $\text{Deconv}_{23}$ layer to $14\times12$ (resolution of the final convolution layer). To do so, we use max and average pooling operations. Since max pooling increases the spatial support of the region, we use it for the masks associated with eyes, eyebrows, nose and mouth. This helps us to capture some context as well. We use average pooling for the remaining regions. For SSG experiments, we use the output of $\text{Deconv}_{33}$ layer, in the semantic segmentation network, as the localization cue. The attribute prediction and semantic segmentation networks are respectively trained for 40K and 75K iterations.

## 6. Conclusion

Aligned with the trend of part-based attribute prediction methods, we proposed employing semantic segmentation to improve facial attribute prediction. Specifically, we transfer localization cues from the auxiliary task of semantic face parsing to the facial attribute prediction problem. In order 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 activation maps to regions that are semantically consistent, SSG carries the same notion but applies it to the earlier layers. We evaluated our proposed methods on CelebA and LFWA datasets and achieved state-of-the-art performance. We also showed that facial attributes can improve semantic face parsing. We hope that this work encourages future research efforts to invest more in the interrelatedness of these two problems.

**Acknowledgments:** We thank anonymous reviewers for insightful feedback, and Amir Emad, Shervin Ardeshir and Shayan Modiri Assari for fruitful discussions. Mahdi M. Kalayeh and Mubarak Shah are partially supported by NIJ W911NF-14-1-0294. Boqing Gong is supported in part by NSF IIS #1566511 and thanks Adobe Systems for a gift.

## References


[1] 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*, pages 955–962, 2013. 2

[2] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 3

[3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *2011 International Conference on Computer Vision*, pages 1543–1550. IEEE, 2011. 2

[4] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 8

[5] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2470–2478, 2015. 2

[6] 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*, pages 346–361. Springer, 2014. 6

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 3, 4

[8] 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*, pages 5375–5384, 2016. 5, 6, 7

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3, 4, 6

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6

[12] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *European conference on computer vision*, pages 340–353. Springer, 2008. 5, 6, 7

[13] 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*, pages 365–372. IEEE, 2009. 2

[14] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012. 5, 7

[15] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2, 3

[16] 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. 2, 5, 6, 7

[17] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 2, 4

[18] 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. 5, 6, 7

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[20] K. K. Singh and Y. J. Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016. 2

[21] 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*, pages 3484–3491, 2013. 2, 4, 5, 7

[22] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2304, 2016. 5, 6, 7

[23] 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*, pages 1637–1644, 2014. 2, 5, 6, 7

[24] 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*, pages 3239–3243. IEEE, 2016. 6, 7