Supplemental Material for Semantic Alignment: Finding Semantically Consistent Ground-truth for Facial Landmark Detection

1. Evaluation Analysis

Although our Semantic Alignment brings significant performance improvement on different datasets. We still use the human annotation as our ground-truth to measure the performance. As explained in our paper, this evaluation might not be precise because the ground-truth also contains annotation noises. In this section, we analysis this problem in two aspects: (1) theoretical analysis and (2) experimental verification.

For (1), We provide the theoretical analysis to support our evaluations. We denote the 'real' ground-truth as $\hat{\mathbf{y}}$ on test set, the human annotation on test set is $\hat{\mathbf{y}}+\epsilon_{tst}$, the noise ϵ_{tst} is assumed as a Gaussian distribution $\epsilon_{tst} \sim \mathcal{N}(0,\, \sigma^2)$. At the same time, we denote f as the landmark detection model trained on training data which has label noise ϵ_{trn}. Then the expected error during test can be formulated as:

$$\begin{aligned} & E\left[(f(\mathbf{x})-\hat{\mathbf{y}}-\epsilon_{tst})^2\right] \\ &= E\left[(f(\mathbf{x})-\hat{\mathbf{y}})^2\right] + E(\epsilon_{tst}^2) - 2E\left[(f(\mathbf{x})-\hat{\mathbf{y}})\epsilon_{tst}\right] \\ &= E\left[(f(\mathbf{x})-\hat{\mathbf{y}})^2\right] + E(\epsilon_{tst}^2) - 2E(f(\mathbf{x})-\hat{\mathbf{y}})E(\epsilon_{tst}) \\ &= E\left[(f(\mathbf{x})-\hat{\mathbf{y}})^2\right] + \sigma^2 \end{aligned} \tag{1}$$

where E represents the expectation, and $f(\mathbf{x})$ is the CNN prediction. The second equality is valid since $f(\mathbf{x})-\hat{\mathbf{y}}$ only depends on the noise ϵ_{trn} acting on the training labels which is assumed to be independent of ϵ_{tst}. From the above equations, we can see that the test set containing human annotation noises (our evaluations) does not affect our conclusions/comparisons because the ideal expected error $E\left[(f(\mathbf{x})-\hat{\mathbf{y}})^2\right]$ and the error with noise $E\left[(f(\mathbf{x})-\hat{\mathbf{y}}-\epsilon_{tst})^2\right]$ have a *constant* difference σ^2.

For(2), in this work, we proposed the 'semantic ambiguity' which means some landmarks (e.g. those evenly distributed along the face contour) do not have consistent and accurate human annotations. It is actually assumed that those 'semantic ambigous' human annotations are all *on* the contour, but are ambigous about the accurate positions on the contour. To answer **Q2**, we introduce a new metric for performance evaluations. Specifically, we use the 'point to line' distance to replace the traditional 'point to point' distance, as shown in Fig. 1. In this way, we can reduce the impact of semantic ambiguity in human ground-truth and better demonstrate our improvements. In Tab. 1, our method gets more significant improvement under the new metric (18.8% vs 13.3%), showing the effectiveness of our method.

Table 1. The results under different evaluation metric.

Metric	HGs(300W FULL%)	HGs + SA(300W FULL%)
point to point	5.04	**4.37(13.3%↑)**
point to line	3.24	**2.63(18.8%↑)**

2. Qualitative Results

Figure 2 and Figure 3 show some qualitative results of the base hourglass detector (HGs) and our Semantic Alignment detector (HGs + SA) on 300W challenging set. It is observed that our method achieves great results for a wide variety of poses, qualities and occlusions. Meanwhile, despite the promising enough overall structures predicted by HGs, the details of HGs + SA are visually superior than the one of HGs (more accurate and regular). Specifically, subfigure (2), (3), (7), (12), (13), (31), (34) show that HGs fails to detect eye landmarks accurately, while our HGs + SA does not, subfigure (3), (5), (6), (7), (11) show that HGs fails to detect contour landmarks accurately, while our HGs + SA does not.

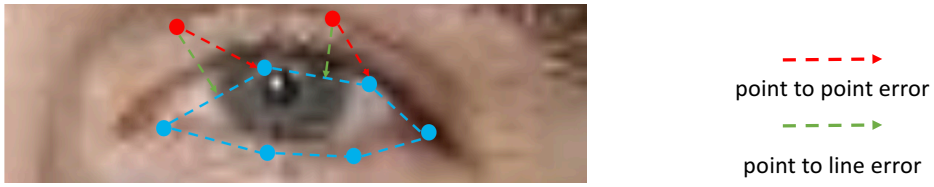

Figure 1. Illustration of point to line error, red dots and blue dots represent the predicted points and human annotations. The error is the distance between the predicted point and target boundary connecting human annotation and its two adjacent points.

Figure 2. The first part of qualitative results on 300-W challenging set. Red and green dots denote the results of baseline (HGs) and Semantic Alignment (HGs + SA), respectively.

References

Figure 3. The second part of qualitative results on 300-W challenging set. Red and green dots denote the results of baseline (HGs) and Semantic Alignment (HGs + SA), respectively.