Delving Deep into Coarse-to-fine Framework for Facial Landmark Localization

Xi Chen, Erjin Zhou, Yuchen Mo, Jiancheng Liu, Zhimin Cao
Megvii Research
{chenxi, zej, moyuchen, liujiancheng, czm}@megvii.com

Abstract

In this paper we proposed a 4-stage coarse-to-fine framework to tackle the facial landmark localization problem in-the-wild. In our system, we first predict the landmark key points on a coarse level of granularity, which sets a good initialization for the whole framework. Then we group the key points into several components and refine each component with local patches cropped within them. After that we further refine them separately. Each key point is further refined with multi-scale local patches cropped according to its nearest 3-, 5-, and 7-neighbors respectively. The results are fused by an attention gate network. Since a different key-point configuration is adopted in our labeled dataset, a linear transformation is finally learned with the least square approximation to adapt our predictions to the competition's task.

1. Introduction

In recent years, significant progress on facial landmark localization has been achieved, especially since Sun *et al*. [24] first applied deep convolutional neural networks (DCNNs) to this problem. Afterwards, tremendous works emerge to improve the performance, which mainly falls into three trends: 1) propose novel network structure to exploit the training data for better generalization ability [25, 4, 34]; 2) modify landmark prediction pipelines with multi-stage strategy [36, 24, 38, 40, 8, 32, 41]; 3) incorporating more information in the training process via, for instance, transfer learning [20].

In this paper, we focus on the second and third methods. In particular, to the best our knowledge, we are the first to raise the transfer problem between tasks with different points, and present a comprehensive framework to tackle it. Our pipeline architecture is generally based on the system proposed by Zhou *et al*. [40], in which they proposed a coarse-to-fine network, which models each facial component refinement as an independent task. Besides, Huang *et al*. [8] extend this idea and propose a multi-scale approach to better grasp and utilize the local information to make more accurate landmark prediction.

Inspired by all these, we proposed a coarse-to-fine framework to predict the landmark predictions by concentrating on more and more fine-grained patches of facial key points. Concretely, our framework consists of four sequential stages: 1) First of all, a pre-trained face detector is utilized to locate the target region, followed by a carefully designed CNN to predict the rotation angle of cropped image. Then the cropped image is rotated to a horizon-canonical position and fed into another CNN to predict the coarse landmark; 2) We then separate the landmark into several components and predict each component's associated landmarks respectively. 3) We further refine each point with multi-scale local patches cropped according to its nearest 3-, 5-, and 7- neighbors. 4) Finally, we transfer the 81-point predicitions to the competition's tasks with lease square approximation.

2. Related Work

Facial landmark localization is a classical topic of research in computer vision. With the development of deep learning, various methods are proposed in recent years. Our method is to refine a coarse landmark estimation through cascades, namely the multi-stage strategy [24, 36, 38, 40, 8, 32, 41]. Another method to perform coarse-to-fine estimation is to use branched networks [15]. Other works have been done to explore different ways to localize keypoints, such as considering keypoint localization as a 3D face model fitting problem [10, 42], initializing with head pose predictions [33] and using separate cluster specific networks [29]. Despite the performance of convolutional neural networks, recent works have also tried to use recurrent neural networks(RNNs) [19, 26, 30] in this task.

Related work also exists in advanced network architectures, especially the Xception [1] and the VGG-16 architectures [22]. The Xception assumes that inter-channel and intra-channel correlations can be entirely decoupled. This

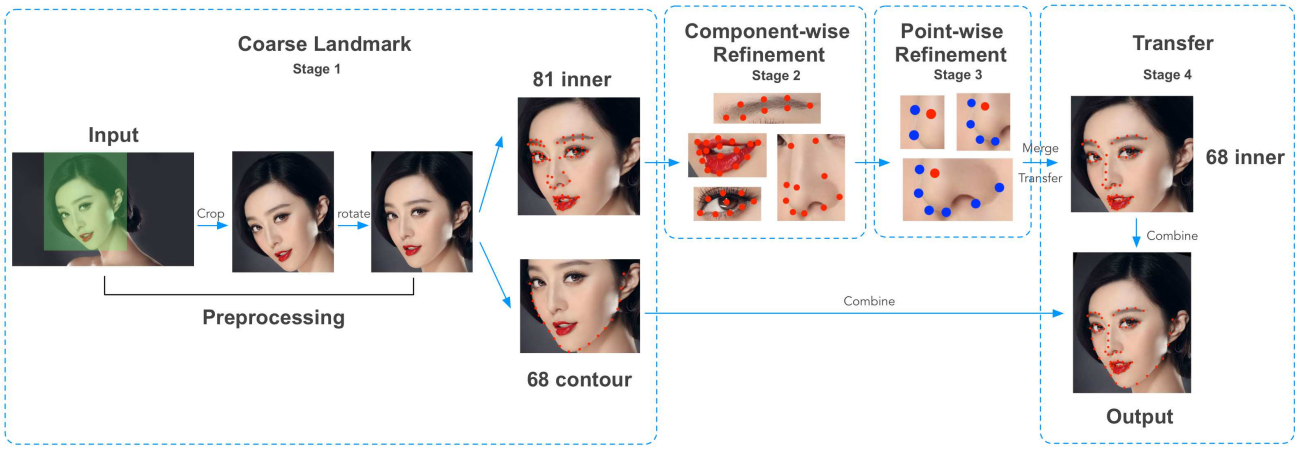


Figure 1. **Framework overview.** Preprocessings like detection and rotation are applied before making any prediction of landmark. Then coarse landmark is produced for both inner and contour points. For inner points, component-wise and point-wise refinements are followed in Stage 2 and Stage 3 respectively. Finally, inner points of the 68-point task are generated from the 81-point task with least square approximation approach, which are combined with contour points of the 68-point task in Stage 1 to obtain the final result.

assumption led to a network structure build with depth-wise (or channel-wise) separable convolution layers [27]. Residual connections, introduced by He *et al*. in [5], are used extensively in our proposed architecture. Furthermore, Xie *et al*. proposed the ResNeXt structure [31] that reintroduced group convolution used in AlexNet [11] into ResNet structure and achieved an improvement in their performance.

Transfer learning build systems that generalize across different domains of different probability distributions [18, 23, 17, 37, 28], which has been widely used in computer vision to achieve better performance on novel domains. The practice of training a CNN on ImageNet [21] and then adapting those features for a new target task was used to solve a wide range of computer vision problems [16, 3]. Experiments and discussions about doing fine-tuning to transfer across domains have been done in [9].

3. Proposed Method

Figure 1 gives a brief illustration of our proposed 4-stage coarse-to-fined framework. Given an input image, we first try to rotate the detected face to the vertical direction, after which we divide all points into two subsets, contour points and inner points. Here we refer contour as point 0 to 18 for the 81-point tasks and point 0 to 16 for the 68-point task. The definition and arrangement of the 81, 68 and 39-point landmark tasks are given in Figure 8. After obtaining coarse landmark from Stage 1, we further separate the inner points into 6 components and refine them with 6 individual networks, which brings a relative 9.75% decrease of the normalized loss. In Stage 3, three different scales of patches are generated for each point and further decrease the normalized loss by 1.77%.

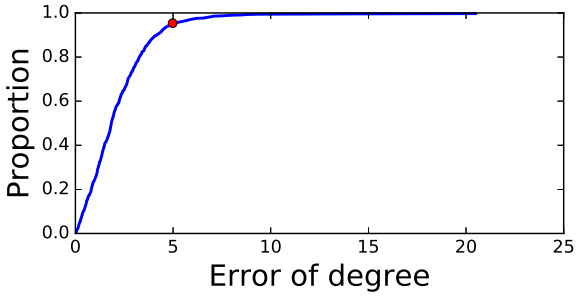


Figure 2. **Performance of angle regression network**

3.1. Coarse Landmark

The Stage 1 is responsible for predicting an initial and coarse landmark, which may not be so accurate but maintains the geometric properties of the whole face.

Generally a facial detector is applied before making any prediction of landmark, which will output bounding box to highlight the major part of face in given image. Then an angle regression network is trained and rotate the face to the vertical direction. These preprocessings are important as they dramatically decrease the location and angle variance of input images. Figure 2 gives the performance of angle regression network and we can see that nearly 95% of images are rotated back to the vertical direction within $\pm 5^\circ$ error.

After rotating the cropped images back to the vertical direction, we divide all points into two subsets, inner points and contour points. We will illustrate the detail reason behind this in Section 4.3. Please keep mind that although we can achieve fairly low normalized loss for both contour and inner points after Stage 3, there are many difficulties when transferring counter points from the 81-point task to the 68-

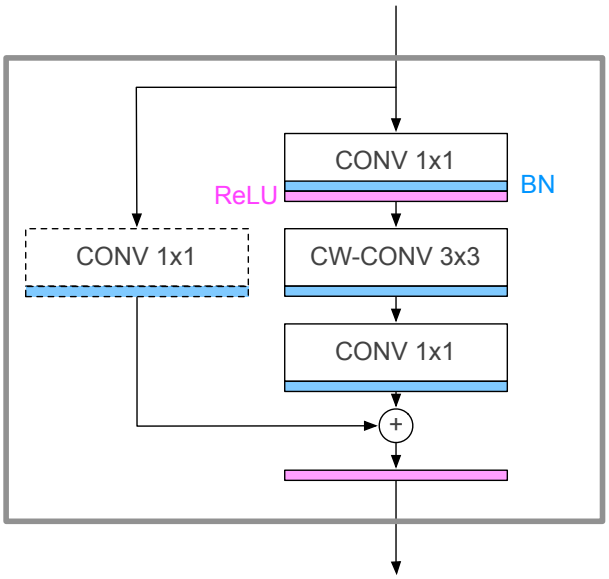


Figure 3. **Base module in Stage 1**

point and the 39-point tasks. Thus we make a trade-off and predict the contour points directly in the 68-point and 39-point tasks during Stage 1.

Inspired by Xception [1] and ResNeXt [31], we use a base module which contains 1x1 convolution, 3x3 channel-wise convolution and skip connection as the building block to construct our network. As shown in Figure 3, 3x3 channel-wise convolution (also known as spatial convolution) is designed for capturing the intra-channel correlation of images, which is surrounded by two 1x1 convolution that learn the inter-channel correlation. Besides, skip connection is important for gradient to propagate back when the network goes deep. We build our coarse landmark prediction network base on this simple model and use teacher-student architecture during training procedure. Figure 4 illustrate the details of our proposed network.

Unlike knowledge distilling proposed by Hinton *et al*. [6] where they use outputs of Softmax as soft label to train student network, we use two identical networks trained with L2 loss. The only difference between teacher and student network is that gradients from loss3 will be stopped before flowing back into teacher network. As a result, student network will receive supervise signal from both ground truth and teacher network. The additional loss term from teacher network expresses as regularization and help student network to avoid overfitting. Surprisingly we find that student network supervised by teacher network in this way consistently outperforms single student network.

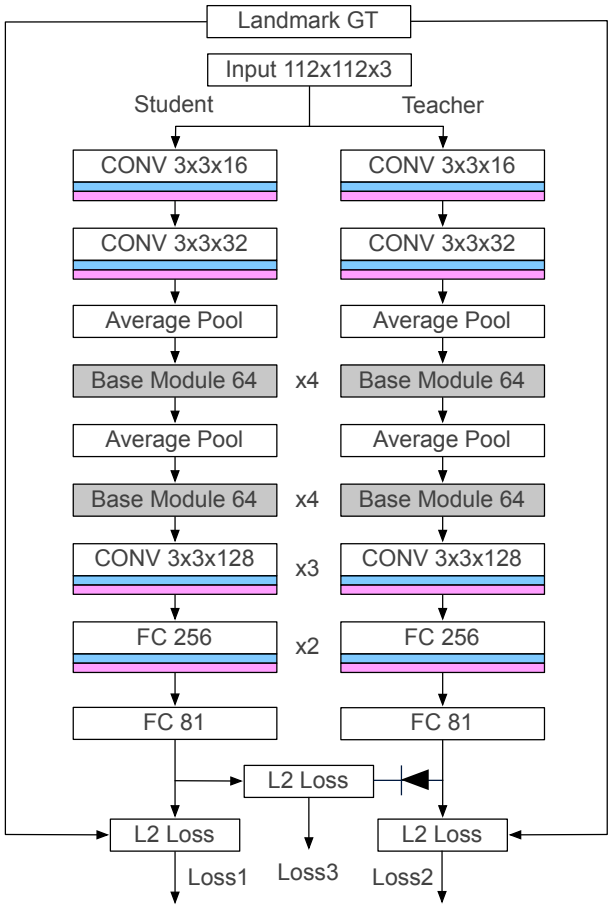


Figure 4. **Network for predicting the coarse landmark in Stage 1** (Gradients flow back to teacher network will be stopped.)

3.2. Component-wise Refinement

Although Stage 1 gives the basic landmark prediction which covers most conditions, it reaches its limitation when the face is either asymmetric or exaggerated. In a typical case, stage 1 network may predict two open eyes when someone closes one eye with another opened. As a result, component-wise refinement is necessary to capture the local variations. In Stage 2, we separate inner points into 6 components, which consists of the left eyebrow, right eyebrow, left eye, right eye, nose, and mouth. For each component, we align the corresponding coarse landmark to the precomputed mean face (see Figure 9), after which we feed the aligned images into the network and predict the relative landmark in the coordinate system of the input image. Thus landmark predictions are meaningful only in the individual coordinate system. They needed to be transformed back to the coordinate system of the original image before delivered into Stage 3.

Besides, for computation efficiency, we use straight-through VGG style network and substitute basic module in

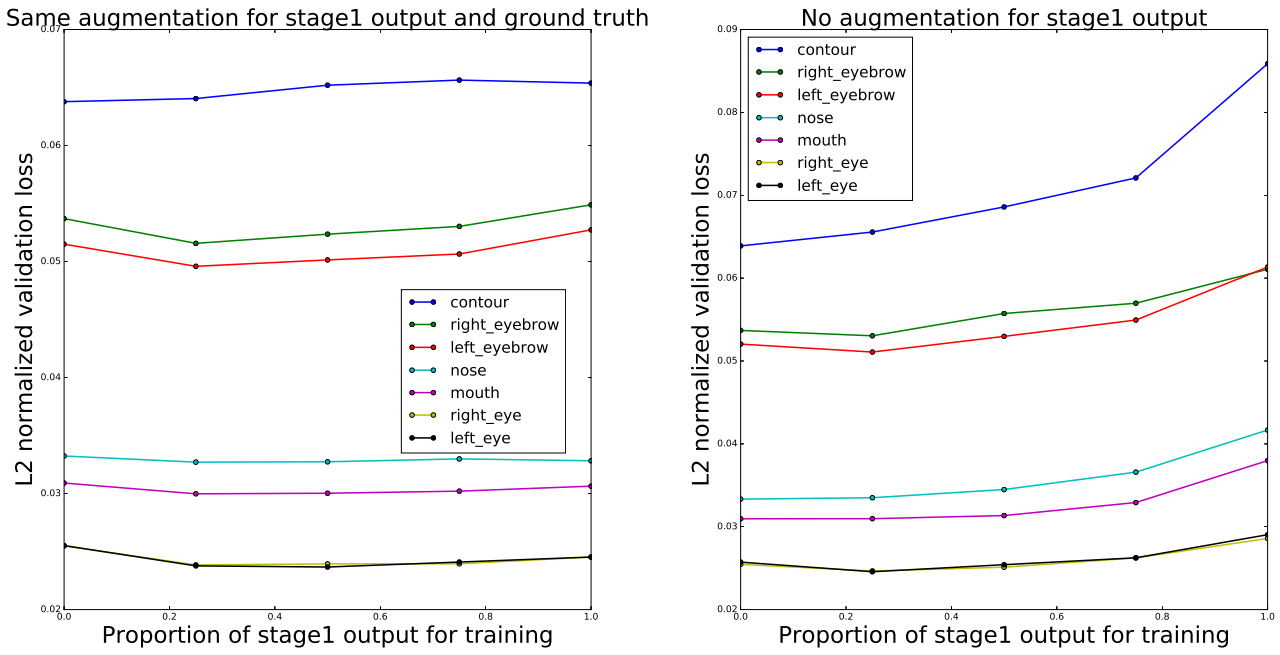


Figure 5. **Effect of different proportion of stage1 output landmark used for training.**

Stage 1 with normal 3x3 convolution layer and eliminate teacher network branch, which sharply decreases the computation complexity from 500M FLOPs to 100M FLOPs.

3.3. Point-wise Refinement

Similar with component-wise refinement, point-wise refinement further refines the prediction landmark in the granularity of each point. On concretely, there are 3 multi-scale patches for each point, which are generated by the nearest 3, 5 and 7 points in mean face respectively. These 3 patches are fed into 3 small networks and predict corresponding 3, 5 and 7 landmark points, with an attentional gate followed by to weight them.

Also, in stage 3 we further reduce the number of channels of all convolution layers and obtain a 25M FLOPs network for each point, which aims to balance the computation complexity of each stage.

3.4. Final transfer

There are several approaches for transferring the P-point landmark task to the Q-point landmark task. One obvious and base method is the least square approximation approach, which tries to minimize the square loss function:

$$\mathbf{T_0} = \arg\min_{\mathbf{T}} ||\mathbf{X} \cdot \mathbf{T} - \mathbf{Y}||_F^2, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{N \times 2P}$, $\mathbf{Y} \in \mathbb{R}^{N \times 2Q}$ and $\mathbf{T} \in \mathbb{R}^{2P \times 2Q}$

Here $\mathbf{X}$ is from source data domain and $\mathbf{Y}$ is from target data domain. $\mathbf{T}$ is a transfer matrix which describes the linear relation between these two domains.

Equation 1 has a close-form solution:

$$\mathbf{T_0} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y}. \tag{2}$$

Besides, we usually expand the source matrix $\mathbf{X}$ by concatenating a column vector $\mathbf{1} \in \mathbb{R}^{N \times 1}$, as well as $\mathbf{Y}$. Then, $\mathbf{T_0}$ has new dimension of $(2P + 1) \times (2Q + 1)$

In our case, P is 81, and Q is either 68 or 39. For simplicity we only introduce the detail of the 68-point task, leaving the 39-point task with a similar treatment. For the 68-point task, we separate all samples into training, validation and test set [35, 26]. We learn the transfer matrix $\mathbf{T_0}$ in training set and apply it to validation and test set. We can evaluate the result on validation set and submit the test set result in the competition.

Alternatively, we can train a neural network to learn the nonlinear relation between the source domain and target domain. However, a worse result is observed compared with the least square approximation approach.

4. Experiments

In this section, we try to investigate the possibility of decoupling of all stages and systematically evaluate the performance of each stage. In addition, we find that contour loss plays an important role when transferring from the 81-point task to the 68-point and 39-point tasks. Moreover, contours loss of tangent direction makes much more contribution to the total loss than that of normal direction.

Landmark Task	Stage1 Coarse Landmark	Stage 2 Component-wise Refinement	Stage 3 Point-wise Refinement	Stage 4 Transfer
81 points (inhouse)	4.317/4.268†	3.896/3.717†	3.827/1.373†	-
68 points (frontal)	-	-	-	4.471‡
39 points (profile)	-	-	-	2.774*

† The format is $v_{\text{real}}/v_{\text{gt}}$ and all results are normalized by distance of two pupils.
‡ Normalized by distance of two pupils.
* Normalized by diagonal distance of tightest bounding box of ground truth landmark.

Table 1. **Performance of each stage.** (All results are multiplied by 100 for brevity)

4.1. Stage Decoupling

The neural network in each stage is trained with images cropped with previous landmark prediction, that makes the whole framework essentially sequential. This brings difficulties in a competition as we want to parallel the training of each stage to shorten the experiment iteration period. Consequently, it is appealing to answer a simple but important question, whether we can decouple each stage and treat them separately.

A natural solution would be to train each stage with the images cropped with ground truth and evaluate with real output of the previous stage. Therefore we conduct an experiment to investigate the difference between coupled approach and this vanilla solution from stage1 to stage2. Here we give the same intensity of augmentation for both stage1 output and ground truth. As shown in Figure 5, we find that too much proportion of stage1 landmark prediction will actually hurt the performance. In particular, except for contour which shows a consistent increment of loss when the proportion varies in $[0.0, 0.25, 0.5, 0.75, 1.0]$, other components have slightly better performance when the proportion reaches 0.25. However, worse performance is obtained when the proportion exceeds 0.25 for all components.

Besides, we conduct another experiment where we only augment those images cropped with ground truth, and find similar tendency with worse performance. As shown in Figure 5, more proportion of non-augmented stage1 output tends to overfit and results in deterioration of performance, which is quite obvious when the proportion reaches 1.0.

As a result, 0.25 may be the 'sweet point' for components except for contour. However, in practice we make a trade-off and utilize decoupled strategy to speed up training procedure.

4.2. Stage Performance

Table 1 illustrates the key results of each stage. Here we evaluate the performance of each stage with not only the real output landmark of previous stage, but also ground truth landmark (First row of Table 1). For clarification we

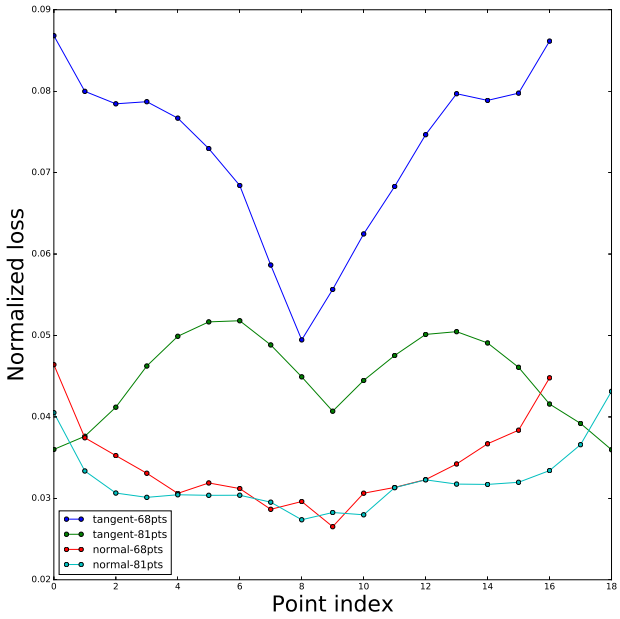
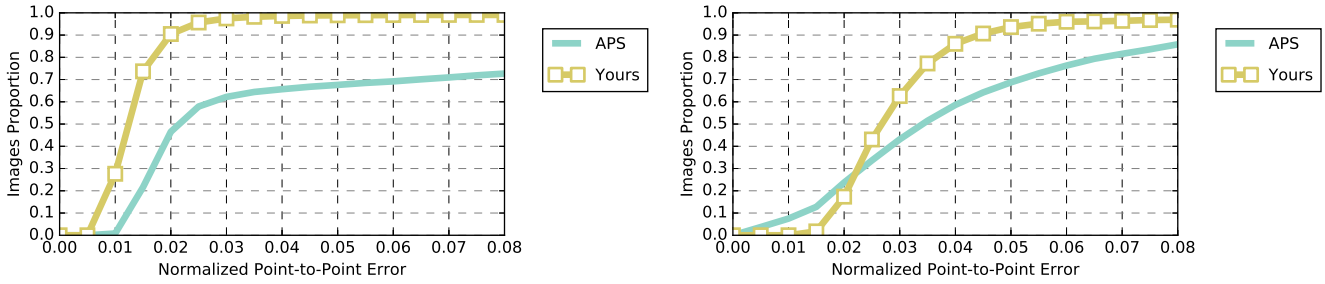


Figure 6. **Tangent and normal loss of contour points**

denote v_{real} and v_{gt} for these two results. We compare v_{real} with v_{gt} and our findings lie in three-fold:

- Both v_{real} and v_{gt} decrease steadily from stage 1 to stage 3, especially for v_{gt} which achieves 1.373 after Stage 3.
- v_{gt} expresses as a lower bound of corresponding stage.
- Except for Stage 3 where v_{real} stands far away from v_{gt}, in Stage 2 and Stage 3 v_{real} is approaching v_{gt}.

The first and second observations are interpretable because we train each stage with ground truth landmark and the network adapt well to the distribution of ground truth landmark. Thus v_{gt} becomes the lower bound of v_{real}. As for the third observation, we can draw a conclusion that component-wise refinement is sufficient for capturing local



(a) Semifrontal (b) Profile

Figure 7. **Performance of test benchmark**

variance and finer granularity helps little for final prediction.

4.3. Transfer Cost

It is unexpected to see that directly transferring points from the 81-point task will result in worse performance. We analyze the loss distribution and find the loss from contour points plays the most important role. Figure 6 further decomposes the contour loss into tangent direction and normal direction and discovering that tangent loss raises sharply when transferring from the 81-point task to the 68-point task. Simply from mean face we can hardly tell the difference of contour points between the 81-point task and the 68-point task (see Figure 8). However, after studying the samples case by case, we seem to find the essential reason.

For the 81-point task, by definition, point 0, 18, 20, 54, 58 and 69 are always stand in the same line. But for the 68-point task, point 0 and 18 have 3D invariant property. This means that no matter what the pose is, point 0 and point 18 refer to the top corner of ear which intersects with the face. As a result, when bowing or raising the head, there will be much inconsistency for the leftmost point between the 81-point and the 68-point tasks, as well as the rightmost point in the contour.

To sum up, in order to predict the leftmost and rightmost key point in the 68-point task, it is necessary to infer the pose of the face with larger receptive field. Thus it is nearly impossible to cover all conditions with a linear transformation matrix $\mathbf{T_0}$. And the raise of the tangent-direction loss in contour points also sheds light on this.

4.4. Public Benchmark Result

We also compare our result with recent state-of-the-arts on the 300-W dataset in Table 2. We have significantly improved the performance and achieve new state-of-the-arts.

4.5. Private Benchmark Result

Figure 7 gives the performance of private test benchmark of semifrontal (the 68-point) and profile (the 39-point) tasks from the organizer [35, 26].

Method	Common	Challenging	Fullset
CFSS [41]	4.73	9.98	5.76
CFSS Practical[41]	4.79	10.92	5.99
cGPRT [14]	-	-	5.71
TCDCN [39]	4.80	8.60	5.54
Fan *et al*. [2]	4.76	8.25	5.45
Honari *et al*. [7]	4.67	8.44	5.41
DCR [12]	4.19	8.42	5.02
Lai *et al*. [13]	4.07	8.29	4.90
Ours	**3.73**	**7.12**	**4.47**

Table 2. **Comparison of state-of-the-arts approaches.**

5. Conclusion

In this paper we present a 4-stage coarse-to-fine framework to address the facial landmark localization problem. We systematically investigate the effectiveness of each stage and successfully decouple them in training procedure. Finally we transfer the inner points of the 81-point task to the 68-point and 39-point ones with least square approximation approach.

References

[1] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016. 1, 3

[2] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016. 6

[3] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011. 2

[4] K. He and X. Xue. Facial landmark localization by part-aware deep convolutional network. In *Pacific Rim Conference on Multimedia*, pages 22–31. Springer, 2016. 1

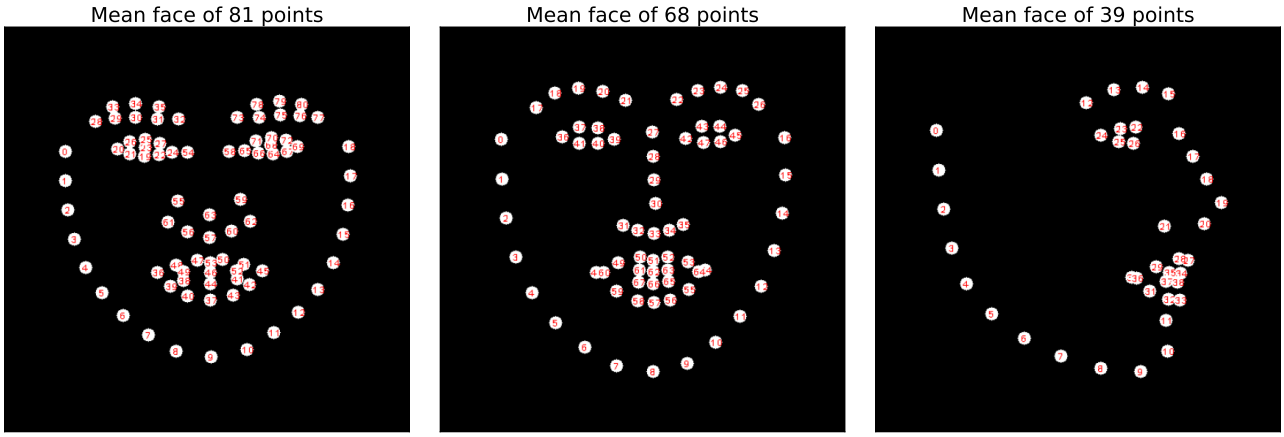


Figure 8. **The definition and arrangement of the 81, 68 and 39-point tasks in mean face.**

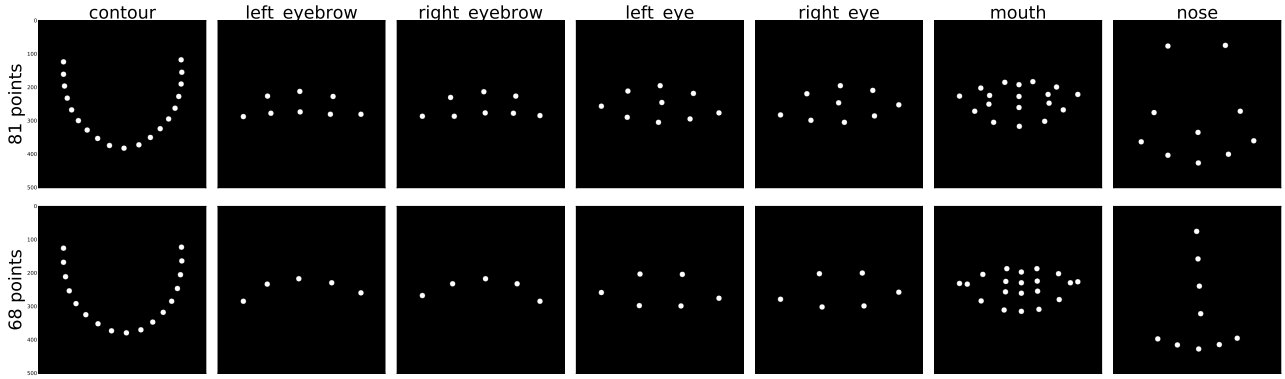


Figure 9. **Mean face of components in the 81-point and 68-point tasks**

[5] 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*, pages 770–778, 2016. 2

[6] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[7] S. Honari, J. Yosinski, P. Vincent, and C. Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[8] Z. Huang, E. Zhou, and Z. Cao. Coarse-to-fine face alignment with multi-scale local patch regression. *arXiv preprint arXiv:1511.04901*, 2015. 1

[9] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2

[10] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016. 1

[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. 2

[12] H. Lai, S. Xiao, Z. Cui, Y. Pan, C. Xu, and S. Yan. Deep cascaded regression for face alignment. *ArXiv e-prints*, 2015. 6

[13] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 6

[14] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4212, 2015. 6

[15] Z. Liang, S. Ding, and L. Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv preprint arXiv:1507.03409*, 2015. 1

[16] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 2

[17] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transac-*

tions on Neural Networks, 22(2):199–210, 2011. 2

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

[19] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *European Conference on Computer Vision*, pages 38–56. Springer, 2016. 1

[20] M. Rashid, X. Gu, and Y. J. Lee. Interspecies knowledge transfer for facial keypoint detection. *arXiv preprint arXiv:1704.04023*, 2017. 1

[21] 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*, 115(3):211–252, 2015. 2

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[23] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008. 2

[24] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013. 1

[25] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. *arXiv preprint arXiv:1705.02193*, 2017. 1

[26] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016. 1, 4, 6

[27] V. Vanhoucke. Learning visual representations at scale, 2014. 2

[28] X. Wang and J. Schneider. Flexible transfer learning under support and model shift. In *Advances in Neural Information Processing Systems*, pages 1898–1906, 2014. 2

[29] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial landmark detection with tweaked convolutional neural networks. *arXiv preprint arXiv:1511.04031*, 2015. 1

[30] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision*, pages 57–72. Springer, 2016. 1

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 2, 3

[32] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 1

[33] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. *arXiv preprint arXiv:1507.03148*, 2015. 1

[34] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision*, pages 52–70. Springer, 2016. 1

[35] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step closer to the solution. In *In IEEE CVPR Workshop on Wide Face*, July 2017. 4, 6

[36] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014. 1

[37] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML (3)*, pages 819–827, 2013. 2

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1

[39] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. 6

[40] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013. 1

[41] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 1, 6

[42] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 1