清華大學
Tsinghua University

i-VisionGroup

Paper Reading

2020-06-18

Fine-Grained Head Pose Estimation Without Keypoints

- Nataniel Ruiz, Eunji Chong, James M. Rehg
- Georgia Institute of Technology

- CVPRW 2018

问题描述

- 头部位姿估计在许多应用中是一个重要问题，如协助视线检测、人脸对齐等。
- 头部位姿估计问题可以看作输入一张2D人脸图像，估计其三维欧拉角的问题。
- 传统的算法是通过关键点检测，并将关键点与一个标准人脸模型做2D-3D点的配准得到。
- 显然，这样的算法受到关键点检测的性能影响，还会受到3D head model的影响，同时还需要一个专门的2d-3d匹配算法。
- 如何设计一种e2e的算法直接计算？

相关工作

- 1、姿态估计作为多任务的子任务：
  - KEPLER：present a modified GoogleNet architecture which predicts facial keypoints and pose jointly.
  - Hyperface：R-CNN+AlexNet,detect faces,determine gender, find landmarks and estimate head pose at once.
  - All-In-One:adds smile, age estimation and facial recognition to the former prediction tasks.
- 2、Landmark-free：
  - Faceposenet: regress 3D head pose using a simple CNN and focus on facial alignment using the predicted head pose.
  - Dynamic:uses a VGG network to regress the head pose Euler angles + a recurrent neural network to improve pose prediction by leveraging the time dimension

1.A.Kumar, A. Alavi, and R. Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 258–265. IEEE, 2017.

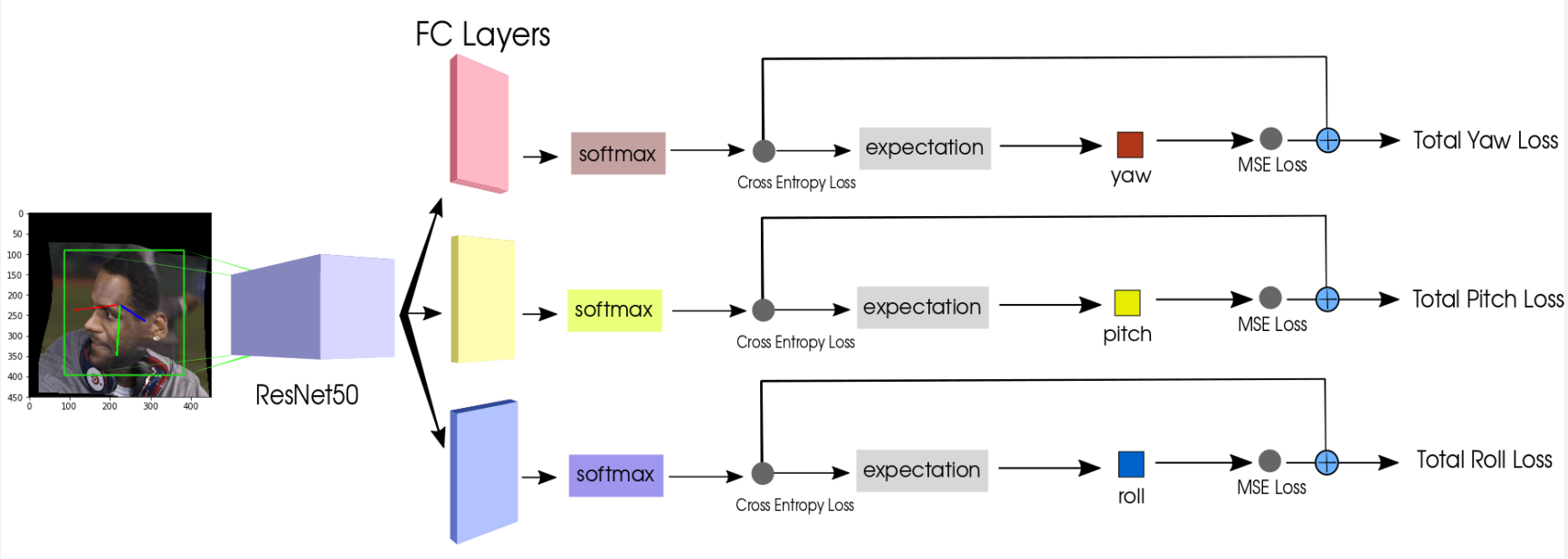
2.R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv preprint arXiv:1603.01249, 2016.

3. R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 17–24. IEEE, 2017.

4.F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Faceposenet: Making a case for landmark-free face alignment. In Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on, pages 1599–1608. IEEE, 2017.

5. J. G. X. Y. S. De and M. J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. 2017.

网络结构


FC Layers
softmax
expectation
Cross Entropy Loss
yaw
MSE Loss
Total Yaw Loss
softmax
expectation
Cross Entropy Loss
pitch
MSE Loss
Total Pitch Loss
softmax
expectation
Cross Entropy Loss
roll
MSE Loss
Total Roll Loss
ResNet50

算法说明

- 使用E2E深度网络的优势：
  - 独立与人脸模型的选取、关键点检测算法的精度、匹配算法的精度；
  - 作为独立估计网络，不会在关键点检测算法出问题时失效；
  - 效果能够逼近使用深度信息的精确算法。
- Loss设计：
  - 直接作为回归问题使用MSELoss并不能达到最好效果；
  - 三个欧拉角，每个单独有一个Loss；
  - 回归问题转分类Loss，把-99到99度每3度做一个角度基点分为66类。
  - 通过期望算法计算最终角度。

$$\mathcal{L} = H(y, \hat{y}) + \alpha \cdot MSE(y, \hat{y})$$

数据库

- 训练库：利用300W-LP数据库生成一批数据。
  - 61k人脸图像和对应的人脸模型。将人脸模型在空间中转动将人脸图像进行变换得到新的图像和对应的三维欧拉角度值。
  - 为解决低分辨率问题，做了随机升降采样和模糊化的数据增广。

- 测试库：AFLW2000和BIWI
  - AFLW2000是AFLW的一个子库，包含了2000个不同人的人脸图像，每个图像有标注68个3D关键点和一个3D模型，可以计算得到人脸图像的姿态。
  - BIWI是在实验室中通过RGBD相机得到的人脸视频，包含大约15,000 帧人脸图像.使用1个人脸模型与点云配准，同时用跟踪算法得到每帧图像的姿态标注。

实验

	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 ($\alpha = 2$)	6.470	6.559	5.436	**6.155**
Multi-Loss ResNet50 ($\alpha = 1$)	6.920	6.637	5.674	6.410
3DDFA [35]	5.400	8.530	8.250	7.393
FAN [2] (12 points)	6.358	12.277	8.714	9.116
Dlib [11] (68 points)	23.153	13.633	10.545	15.777
Ground truth landmarks	5.924	11.756	8.271	8.651

Table 1. Mean average error of Euler angles across different methods on the AFLW2000 dataset [35].

	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 ($\alpha = 2$)	5.167	6.975	3.388	5.177
Multi-Loss ResNet50 ($\alpha = 1$)	4.810	6.606	3.269	**4.895**
KEPLER [14]†	8.084	17.277	16.196	13.852
Multi-Loss ResNet50 ($\alpha = 1$)†	5.785	11.726	8.194	8.568
3DMM+ Online [33] *	2.500	1.500	2.200	**2.066**
FAN [2] (12 points)	8.532	7.483	7.631	7.882
Dlib [11] (68 points)	16.756	13.802	6.190	12.249
3DDFA [35]	36.175	12.252	8.776	19.068

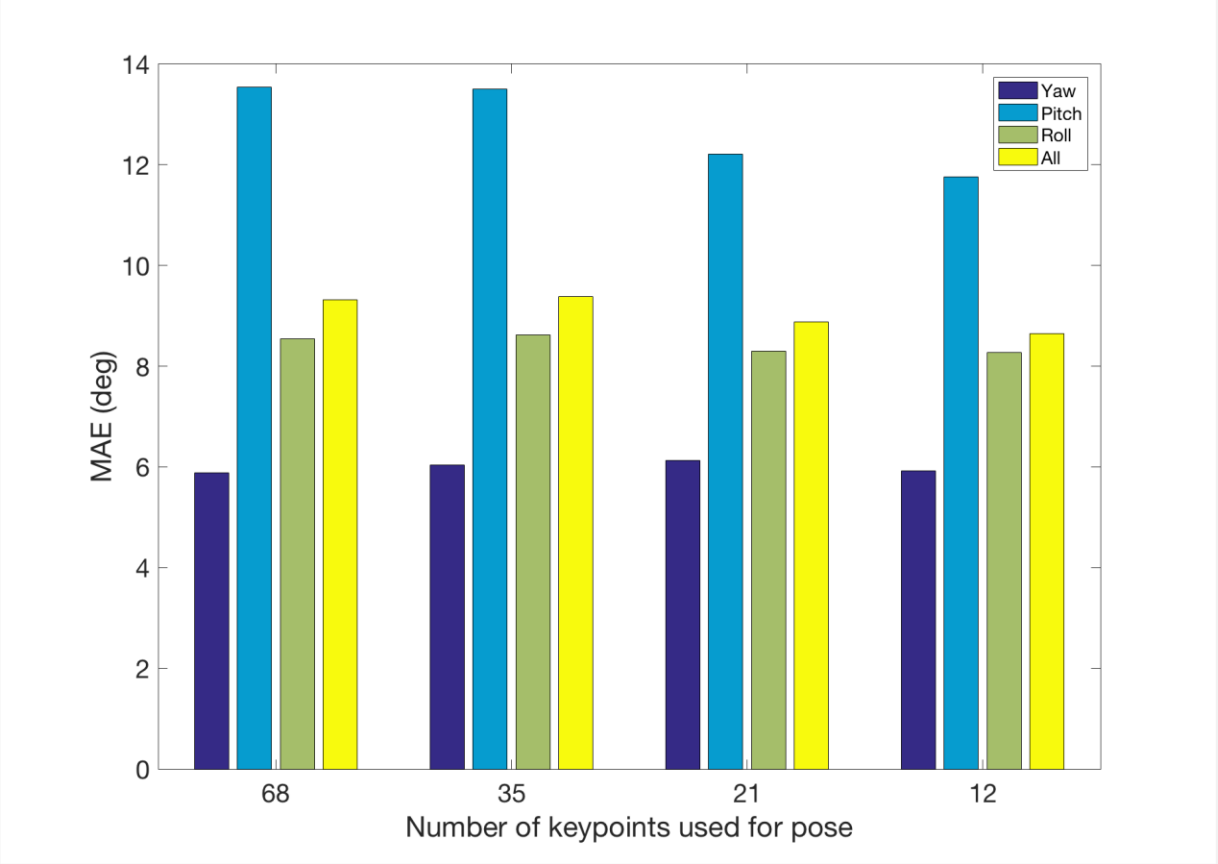
Table 2. Mean average error of Euler angles across different methods on the BIWI dataset [6]. * These methods use depth information. † Trained on AFLW

	Yaw	Pitch	Roll	Sum of errors
Multi-Loss ResNet50 ($\alpha = 1$)	**3.29**	**3.39**	**3.00**	**9.68**
Gu et al. [5]	3.91	4.03	3.03	10.97

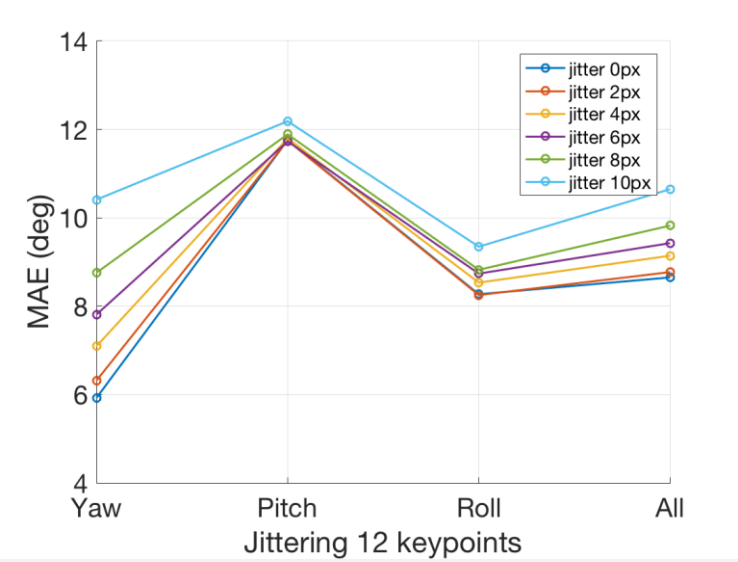
Table 3. Comparison with Gu et al. [5]. Mean average error of Euler angles averaged over train-test splits of the BIWI dataset [6].

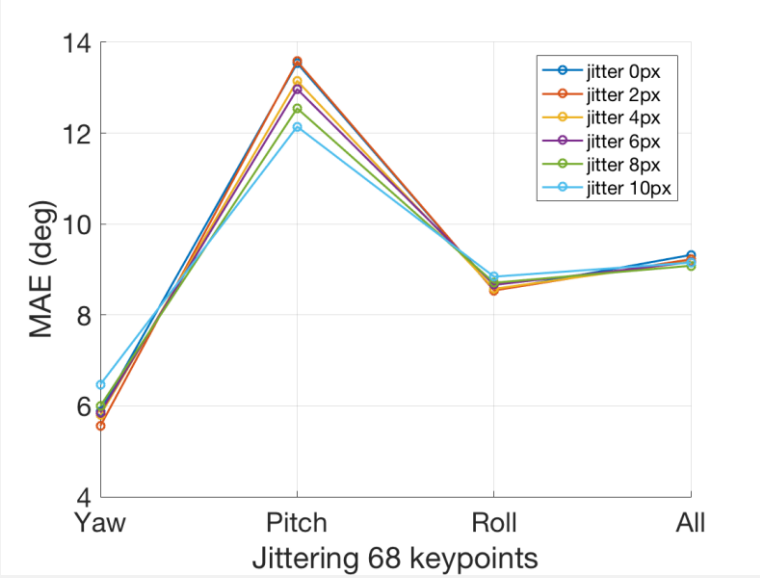
分析基于关键点的姿态估计性能

- 为了验证关键点方案的性能，作者做了以下实验：
- 1、使用GT关键点和GT平均人脸模型，计算使用不同数量关键点的结果

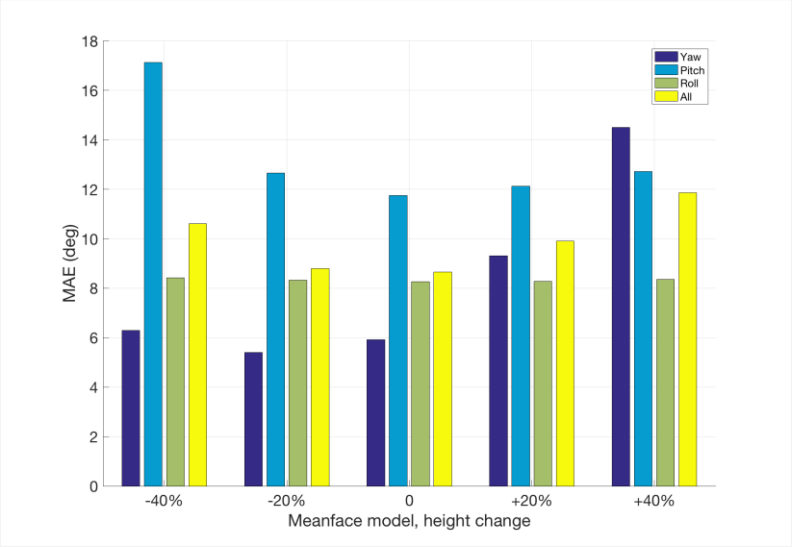


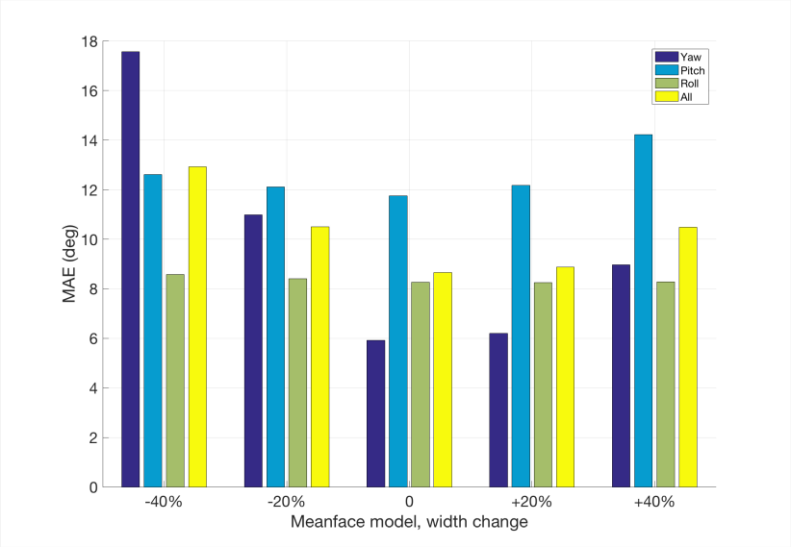
2、使用GT关键点，但做一定量的偏移



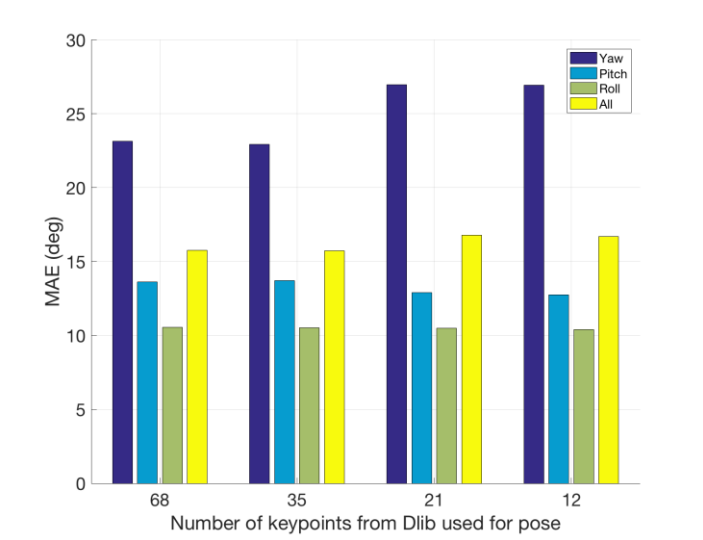


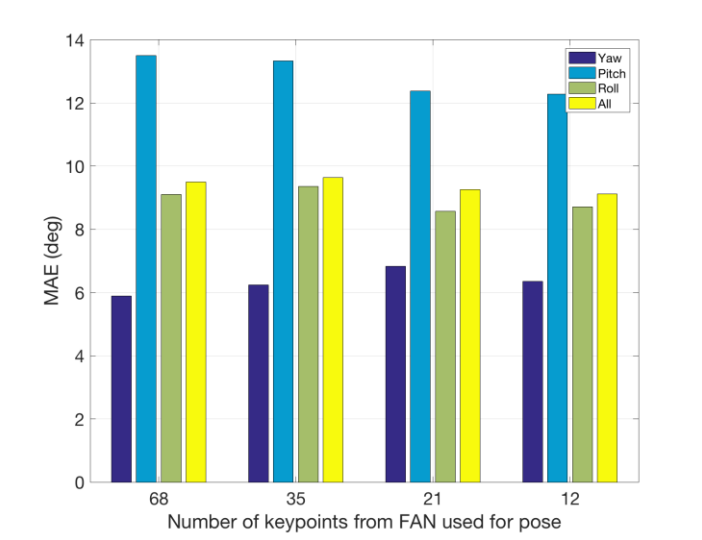
3、改变人脸模型的宽度和高度





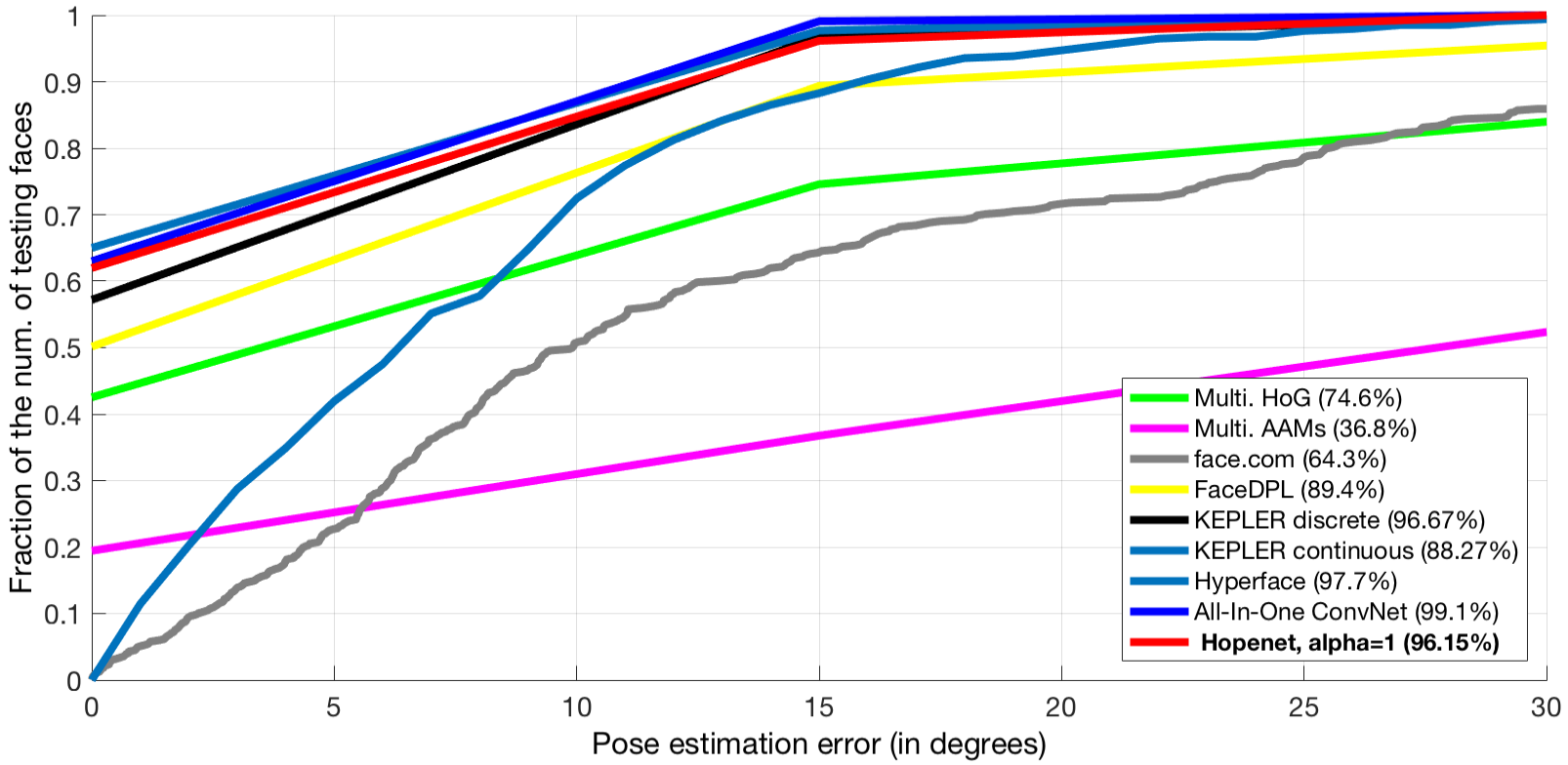
- 4、使用算法得到关键点





- 结论：关键点越准确，需要的点越少。反之亦然。使用算法的话，无法评估这里的tradeoff

AFW性能

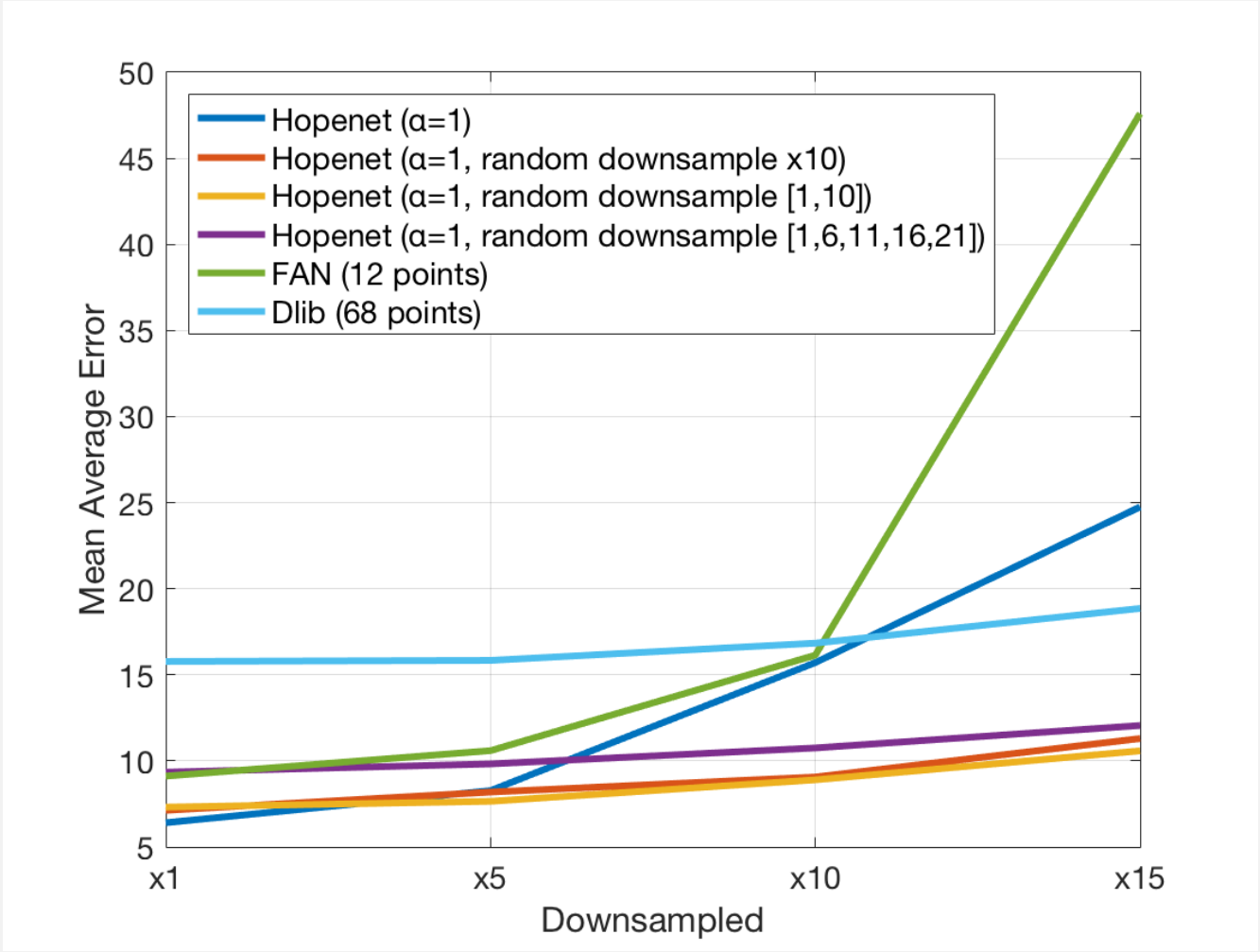

Fraction of the num. of testing faces
Pose estimation error (in degrees)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
Multi. HoG (74.6%)
Multi. AAMs (36.8%)
face.com (64.3%)
FaceDPL (89.4%)
KEPLER discrete (96.67%)
KEPLER continuous (88.27%)
Hyperface (97.7%)
All-In-One ConvNet (99.1%)
Hopenet, alpha=1 (96.15%)

Multi-Loss分析

	α	Yaw	Pitch	Roll	MAE
ResNet50 regression only		13.110	6.726	5.799	8.545
Multi-Loss ResNet50	4	7.087	6.870	5.621	6.526
	2	6.470	6.559	5.436	**6.155**
	1	6.920	6.637	5.674	6.410
	0.1	10.270	6.867	5.420	7.519
	0.01	11.410	6.847	5.836	8.031
	0	11.628	7.119	5.966	8.238
Multi-Loss AlexNet	1	27.650	8.543	8.954	15.049
	0.1	30.110	9.548	9.273	16.310
	0.01	25.090	8.442	8.287	13.940
	0	24.469	8.350	8.353	13.724

Table 5. Ablation analysis: MAE across different models and regression loss weights on the AFLW2000 dataset.

低分辨率分析


Hopenet (α=1)
Hopenet (α=1, random downsample x10)
Hopenet (α=1, random downsample [1,10])
Hopenet (α=1, random downsample [1,6,11,16,21])
FAN (12 points)
Dlib (68 points)
Mean Average Error
50
45
40
35
30
25
20
15
10
5
x1
x5
x10
x15
Downsampled

总结

- 设计了一种E2E的直接估计人脸欧拉角的网络，通过多Loss的设计，比之前的算法都要准确。
- 网络通过一个生成的数据库训练，不需要在其他数据库上FineTune就可以直接使用。
- 分析了不同的backbone网络和不同Loss系数的性能
- 分析了使用关键点进行姿态估计的性能，从而证明本算法的优势
- 对低分辨率图像通过降采样训练数据来处理
- 源码：https://github.com/nataniellruiz/deep-head-pose

FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image

- Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, Yung-Yu Chuang
- Academia Sinica, Taiwan & National Taiwan University, Taiwan
- CVPR2019

算法简述

- 这篇论文提出的FSA-Net，从单张RGB图像就可进行姿态估算，其不需要人脸关键点的提供，直接对人脸的姿态进行回归，我们提出的这种方法使用的是 soft stagewise regression策略。

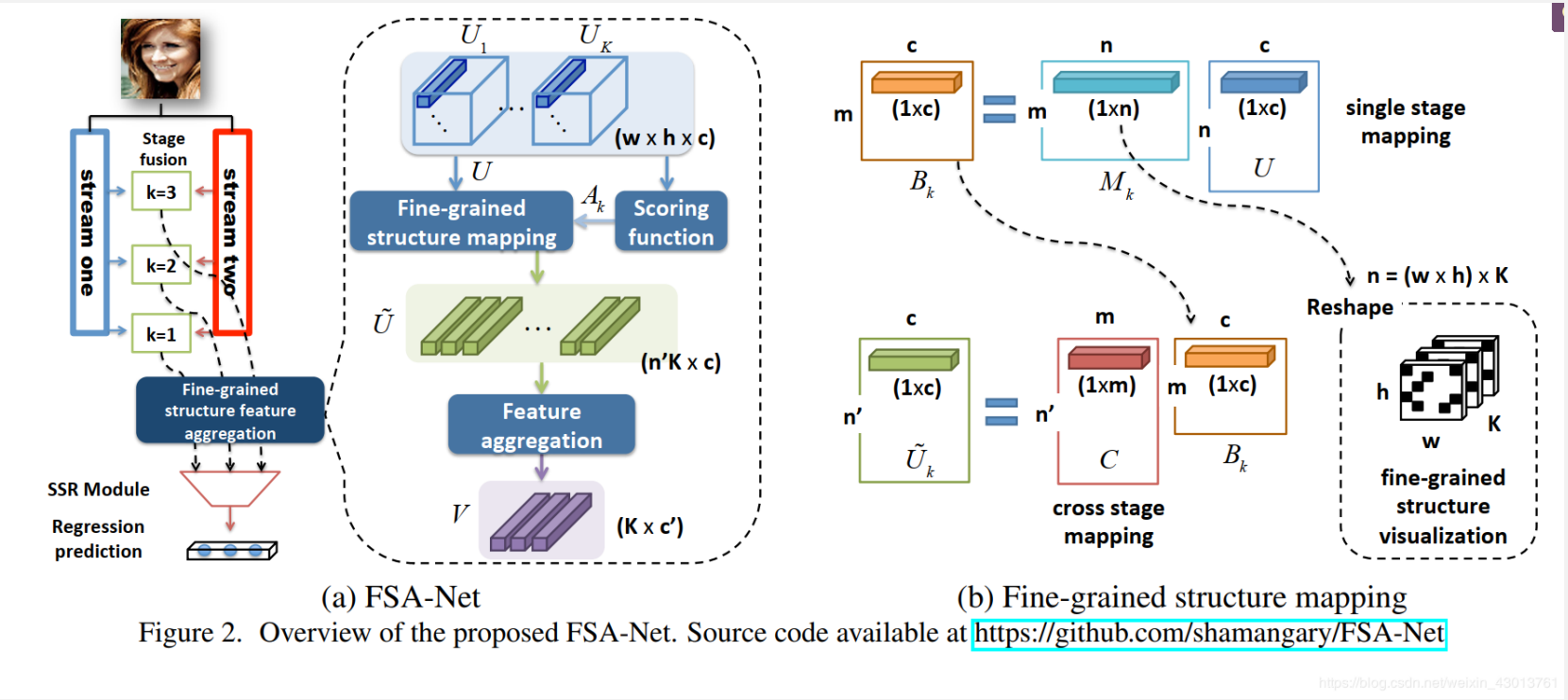
SSR-Net-MD

- 从年龄估计网络SSR-Net出发
- 回归问题转变分类问题
  - 0-99
  - 3类[0，33],[33,66],[66,99]，误差大
  - 9类{0,11},{11,22},{22,33},{33,44}…{88,99}，精度提高，参数也变大。
  - coarse-to-fine：{[0,11],[11,22],[22,33]},
    {[33,44],[44,55],[55,66]}，{[66,77],[77,88],[88,99]}

$$\tilde{y} = \sum_{k=1}^{K} \vec{p}^{(k)} \cdot \vec{\mu}^{(k)},$$

  - 为了更好的去表示这个u，作者又提出了η（中心偏移量）以及Δ（对bins的宽度进行缩放），因此u使用η以及Δ进行代替，网络学习的就是这样的K组三元组 $\{\vec{p}^{(k)}, \vec{\eta}^{(k)}, \tilde{\Delta}_k\}_{k=1}^{K}$

网络结构

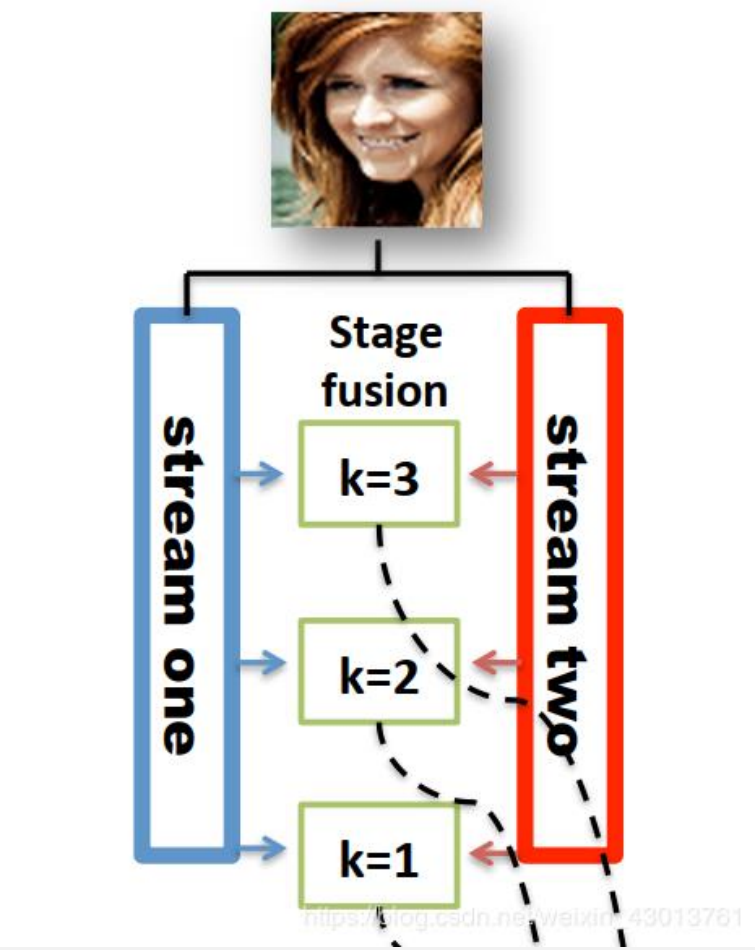


(a) FSA-Net (b) Fine-grained structure mapping

Figure 2. Overview of the proposed FSA-Net. Source code available at https://github.com/shamangary/FSA-Net

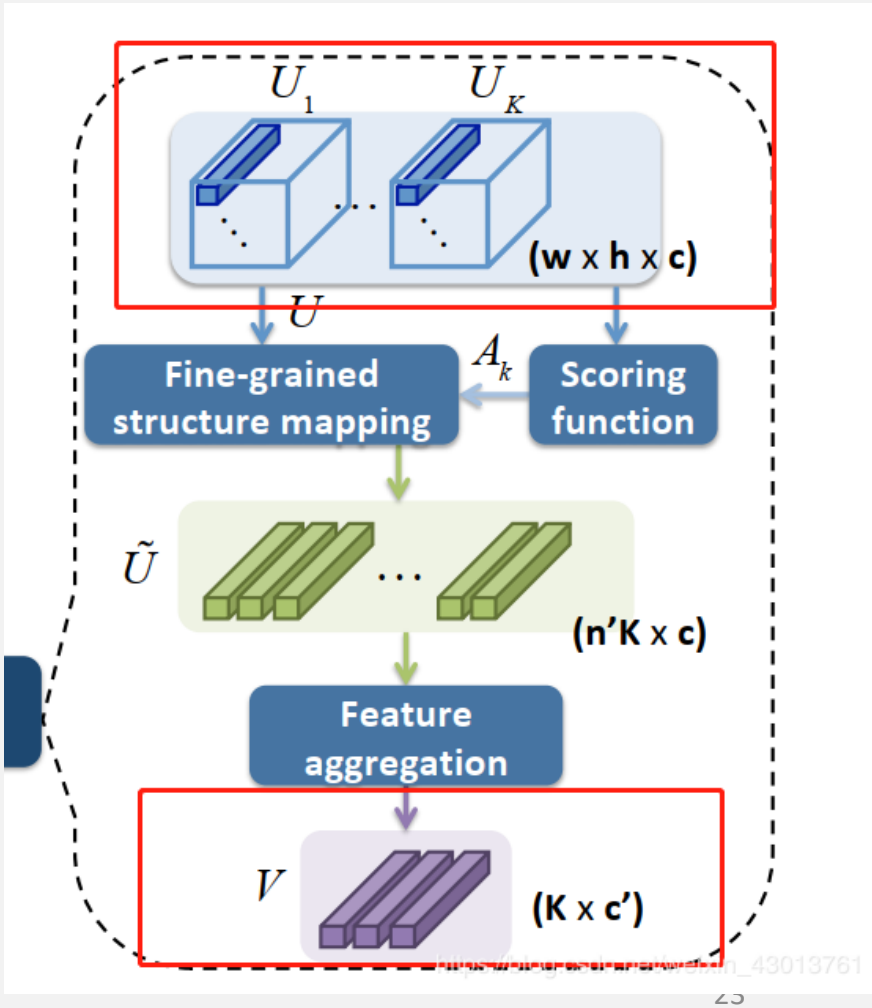
Backbone

- 特征提取部分：
- 输入图片需要经过两个streams，每个stream分别会在K=3，和K=2，以及K=1处分别提出特征，然后两两配对，通过1x1的卷积进行融合，输出C通道的特征图。即每一层会得到一个w×h×c的特征图。



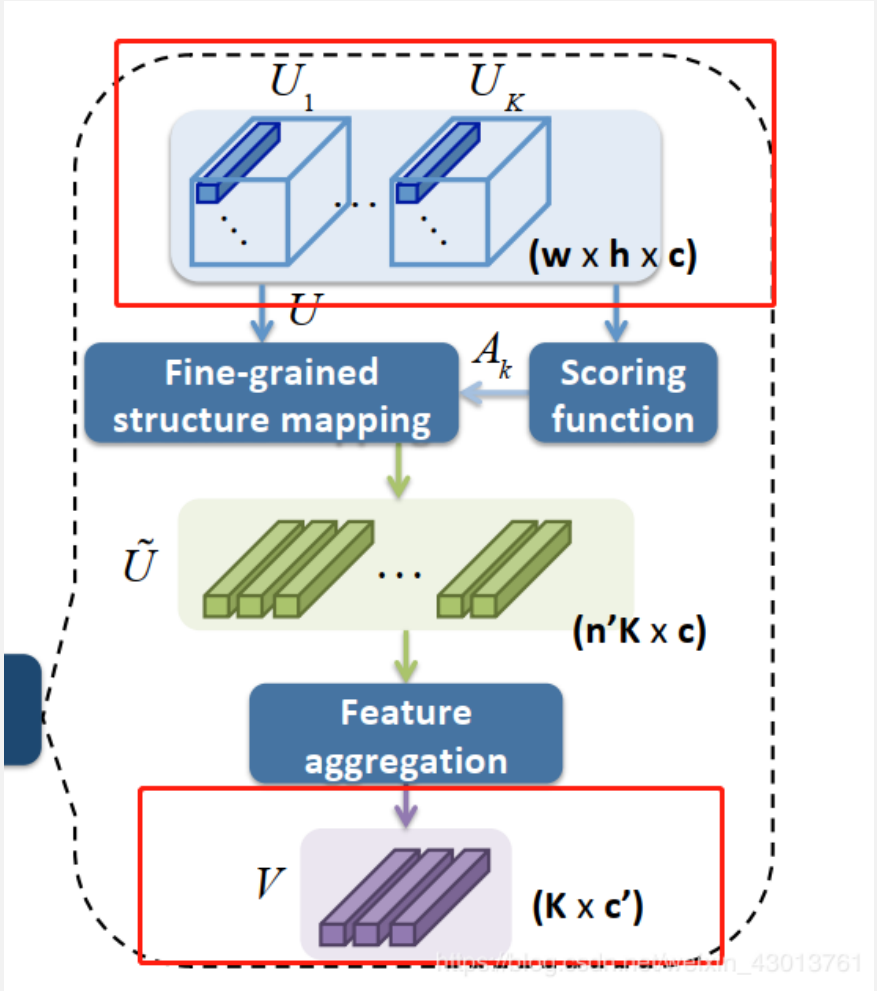
Fine-grained aggregation

- 给出K个w×h×c大小的特征图，aggregation 模型的的任务就是对他们进行提炼，获得更加小的，K个维度为c’的特征向量。最终V会通过SSR模块计算得到姿态结果。



Scoring function

- 加入Scoring function注意力机制Ak，对w×h×c大小的特征图Uk进行像素级别的重要性评分
- 分三种注意力机制：
- 1、1*1卷积（学习，可能过拟合）
- 2、方差（非学习）
- 3、无注意力（全部置1）



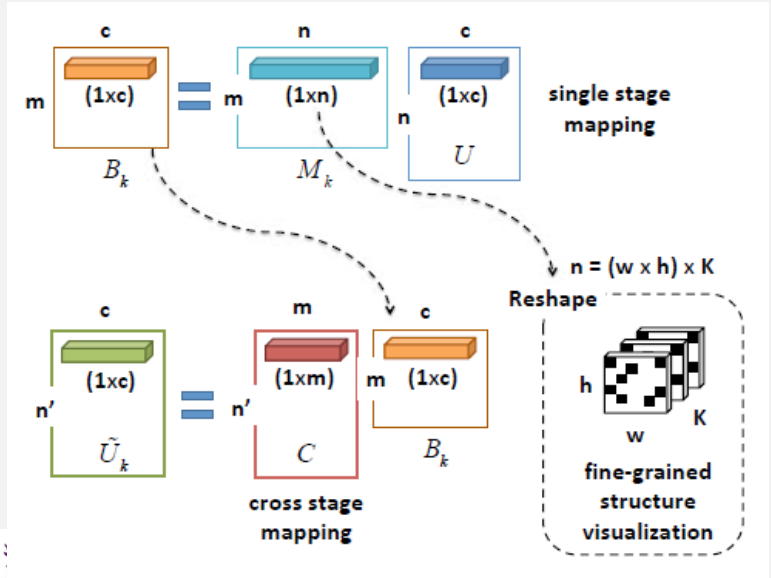
Fine-grained structure mapping

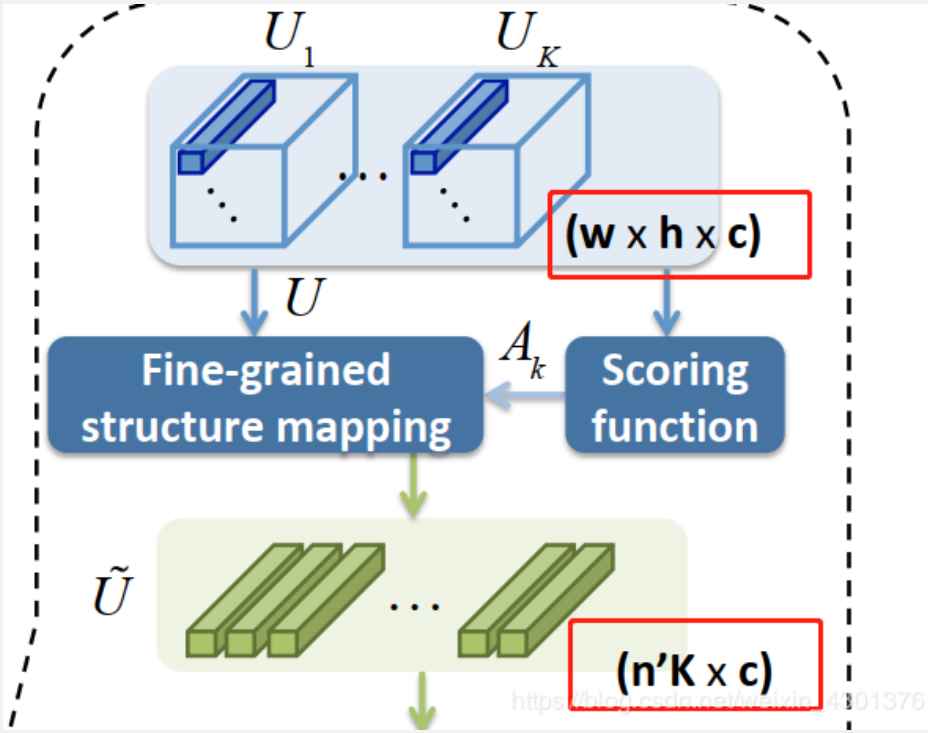
- 有了U和Ak，我们下一步要得到 $\tilde{U}$
- 需要找到一个映射

$$\tilde{U}_k = S_k U,$$

- 我们把Sk分为两部分：

$$S_k = C M_k$$



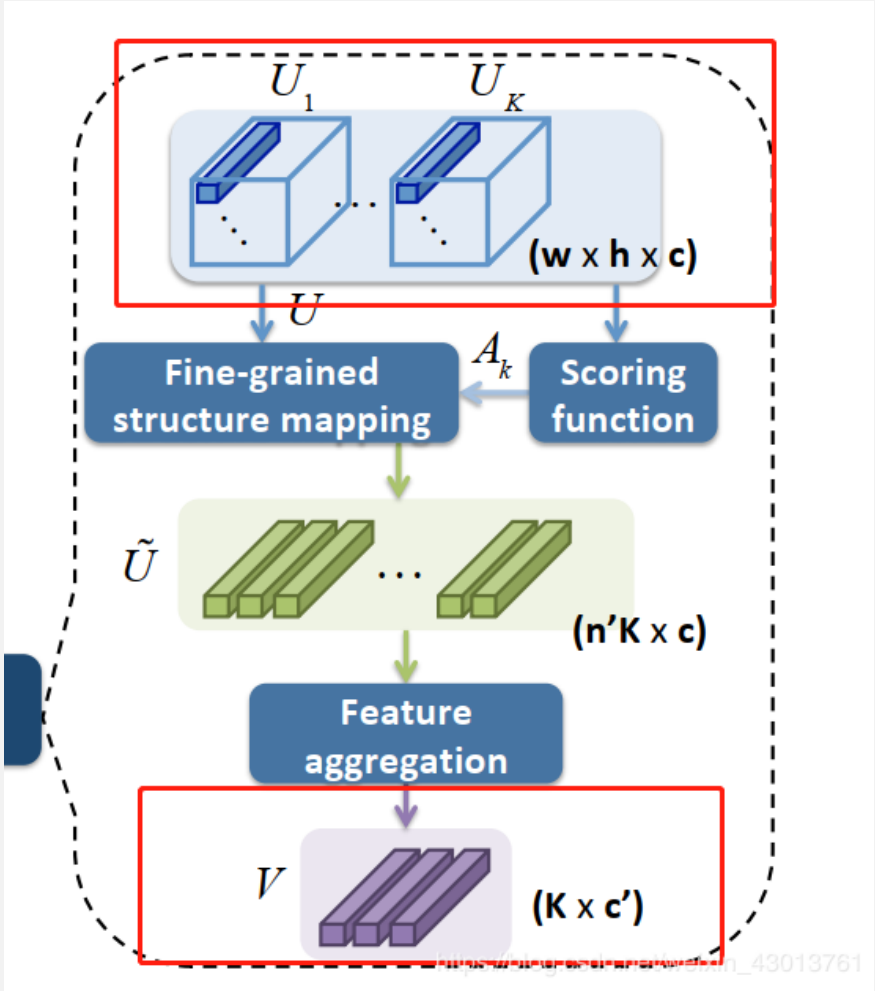


Fine-grained structure mapping

- 其中Mk表示层内的映射，C表示跨层的映射

$$M_k = \sigma(f_M(A_k)),$$
$$C = \sigma(f_C(A)),$$

- 更进一步，我们还对Sk的每一行进行了L1正则化。
- 最终把所有层的Uk~并起来得到U~，输入 aggregation模型*得到V



Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In Proceedings of Neural Information Processing Systems Conference (NIPS), pages 3856–3866, 2017.

网络细节

- 两个streams的建立都是基于BR和BT两个模块

$$\mathrm{B_R}(c) \equiv \{\text{SepConv2D}(3\times 3, c)\text{-BN-ReLU}\},$$
$$\mathrm{B_T}(c) \equiv \{\text{SepConv2D}(3\times 3, c)\text{-BN-Tanh}\},$$

- Stream1:

$$\{B_R(16) - AvgPool(2\times 2) - B_R(32) - B_R(32) - AvgPool(2\times 2)\}$$
$$\{B_R(64) - -B_R(64) - AvgPool(2\times 2)\}$$
$$\{B_R(128) - -B_R(128)\}$$

- Stream2:

$$\{B_T(16) - MaxPool(2\times 2) - B_T(32) - B_T(32) - MaxPool(2\times 2)\}$$
$$\{B_T(64) - -B_T(64) - MaxPool(2\times 2)\}$$
$$\{B_T(128) - -B_T(128)\}$$

- 其他一些参数：$w = 8$, $h = 8$, $c = 64$, $m = 5$, $n' = 7$, $c' = 16$

实验

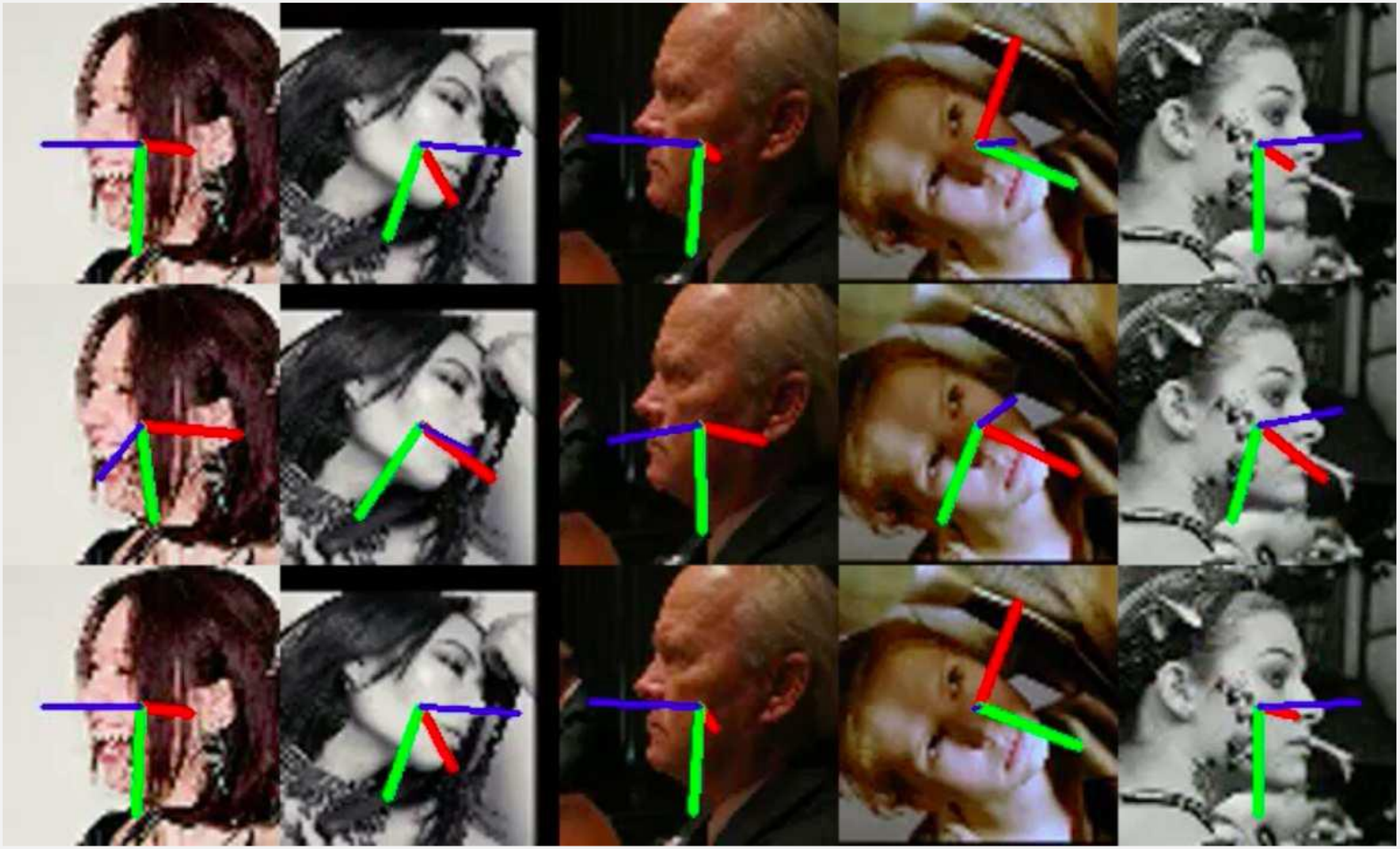
- 数据集：300W-LP训练，AFLW2000和BIWI测试

	MB	Yaw	Pitch	Roll	MAE
Dlib (68 points) [20]	-	23.1	13.6	10.5	15.8
FAN (12 points) [3]	183	6.36	12.3	8.71	9.12
Landmarks [35]	-	5.92	11.86	8.27	8.65
3DDFA [48]	-	5.40	8.53	8.25	7.39
Hopenet ($\alpha=2$) [35]	95.9	6.47	6.56	5.44	6.16
Hopenet ($\alpha=1$) [35]	95.9	6.92	6.64	5.67	6.41
SSR-Net-MD [45]	**1.1**	5.14	7.09	5.89	6.01
FSA-Caps (w/o)	2.9	5.27	6.71	5.28	5.75
FSA-Caps (1×1)	**1.1**	4.82	6.19	4.76	5.25
FSA-Caps (var.)	**1.1**	4.96	6.34	4.78	5.36
FSA-Caps-Fusion	5.1	**4.50**	**6.08**	**4.64**	**5.07**

Table 1. Comparisons with the state-of-the-art methods on the AFLW2000 dataset. All are trained on the 300W-LP dataset.

	MB	Yaw	Pitch	Roll	MAE
3DDFA [48]	-	36.2	12.3	8.78	19.1
KEPLER [22]	-	8.80	17.3	16.2	13.9
Dlib (68 points) [20]	-	16.8	13.8	6.19	12.2
FAN (12 points) [3]	183	8.53	7.48	7.63	7.89
Hopenet ($\alpha=2$) [35]	95.9	5.17	6.98	3.39	5.18
Hopenet ($\alpha=1$) [35]	95.9	4.81	6.61	3.27	4.90
SSR-Net-MD [45]	**1.1**	4.49	6.31	3.61	4.65
FSA-Caps (w/o)	2.9	4.56	5.15	2.94	4.22
FSA-Caps (1×1)	**1.1**	4.78	6.24	3.31	4.31
FSA-Caps (var.)	**1.1**	4.56	5.21	3.07	4.28
FSA-Caps-Fusion	5.1	**4.27**	**4.96**	**2.76**	**4.00**

Table 2. Comparisons with the state-of-the-art methods on the BIWI dataset. All are trained on the 300W-LP dataset.

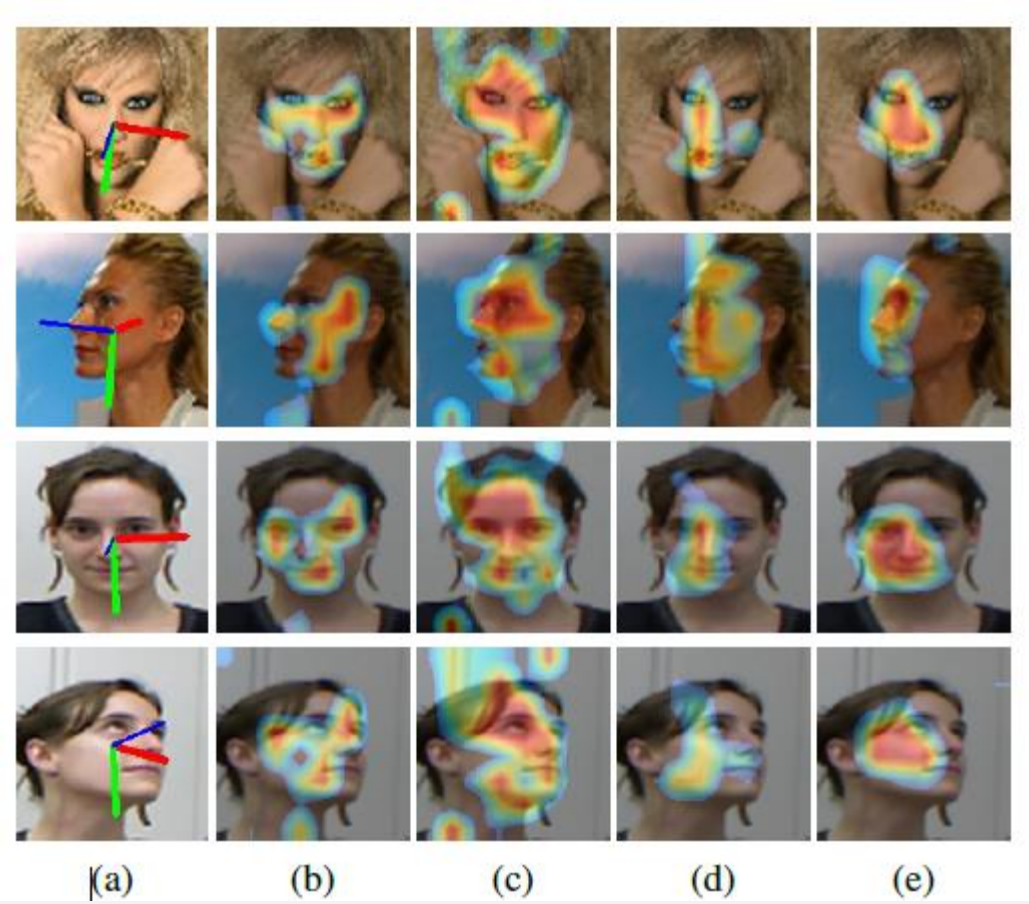


实验2

- 使用BIWI的70%的视频帧做训练，剩余做测试

Method	MB	Yaw	Pitch	Roll	MAE
RGB-based					
DeepHeadPose [28]	-	5.67	5.18	-	-
SSR-Net-MD [45]	**1.1**	4.24	4.35	4.19	4.26
VGG16 [16]	500	3.91	4.03	3.03	3.66
FSA-Caps-Fusion	5.1	**2.89**	4.29	3.60	3.60
RGB+Depth					
DeepHeadPose [28]	-	5.32	4.76	-	-
Martin [25]	-	3.6	**2.5**	**2.6**	2.9
RGB+Time					
VGG16+RNN [16]	>500	3.14	3.48	**2.60**	**3.07**

注意力机制可视化

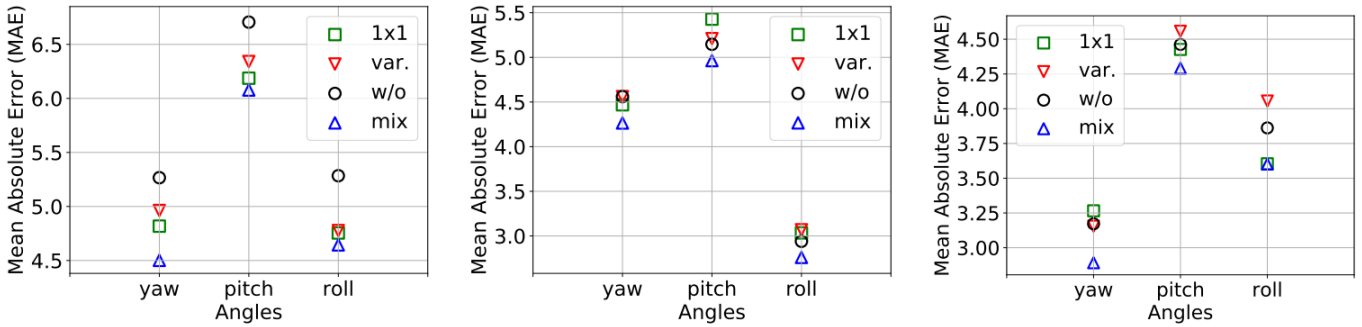


消融学习

testing set	AFLW2000 (protocol 1)							BIWI (protocol 1)						
method	SSR	FSA-Net						SSR	FSA-Net					
aggregation	-	-			Capsule [36]			-	-			Capsule [36]		
pixelwise scoring	-	w/o	1×1	var.	w/o	1×1	var.	-	w/o	1×1	var.	w/o	1×1	var.
model size (MB)	1.1	0.5	0.8	0.8	2.9	1.1	1.1	1.1	0.5	0.8	0.8	2.9	1.1	1.1
MAE	6.01	5.54	5.48	5.41	5.75	5.25	5.36	4.65	4.61	4.53	4.16	4.22	4.31	4.28
MAE (late fusion)	-	5.14			**5.07**			-	4.19			**4.00**		

Table 4. Ablation study for different aggregation methods (no aggregation and Capsule) and the different pixelwise scoring functions for protocol 1. The results are the MAEs of the yaw, pitch, and roll angles. SSR denotes SSR-Net-MD [45].

testing set	BIWI (protocol 2)									
method	SSR-Net-MD [45]	FSA-Net								
aggregation	-	-			Capsule [36]			NetVLAD [1]		
pixelwise scoring	-	w/o	1×1	var.	w/o	1×1	var.	w/o	1×1	var.
model size (MB)	1.1	0.5	0.8	0.8	2.9	1.1	1.1	0.6	0.8	0.8
MAE	4.26	3.95	4.01	3.83	3.84	3.77	3.92	3.97	3.88	3.88
MAE (late fusion)	-	3.75			**3.60**			3.68		



(a) AFLW2000 (protocol 1) (b) BIWI (protocol 1) (c) BIWI (protocol 2)

Figure 6. Comparisons over each angle for different testing datasets and corresponding protocols. We divide the components of FSA-Caps-Fusion into three parts, 1×1, var., and w/o variants. The legend "mix" represents the fusion model.