Supplementary Information for Improving Landmark Localization with Semi-Supervised Learning

S.5.1. Comparison on MTFL dataset

In table S1 we compare with other models on MTFL [49] dataset which provides 5 landmarks on facial images: eye-centers, nose tip, mouth corners. We follow the same protocol as [13] for comparison, where we use train and valid sets of 9,000 and 1,000 images, respectively. We test our model on AFLW and AFW subsets, with 29,995 and 337 images, that were re-annotated with 5 landmarks. For the $L + A$ case we use the head-pose which is categorized into one of the five cases: right profile, right, frontal, left, left profile. Other attribute labels, e.g. gender and wearing glasses, cannot be determined from such few landmarks and therefore are not useful in our proposed semi-supervised learning of landmarks.

Table S1: Results on MTFL test sets for 100% labelled data

	Model					Our	
	ESR	RCPR	SDM	TCDCN	RCN	RCN+(L)	RCN+(L+A)
AFLW	12.4	11.6	8.5	8.0	5.6	5.22	**5.02**
AFW	10.4	9.3	8.8	8.2	5.36	5.13	**5.08**

S.5.2. Selecting auxiliary labels for semi-supervised learning

The impact of an attribute on the landmark in sequential training depends on the amount of informational overlap between the attribute and the landmarks. We suggest to measure the normalized mutual information adjusted to randomness (Adjusted Mutual Information (AMI)), as a selection heuristic, prior to applying our method. AMI ranges from 0 to 1 and indicates the fraction of statistical overlap. We compute for each attribute its AMI with all landmark coordinates.

On Multi-PIE we got AMI(x;y) = 0.045, indicating a low mutual information between coordinates x and y. We therefore compute AMI for attribute (A) and every landmark (as x,y pair) by discretizing every variable uniformly under assumption of coordinate independence: AMI(A;x,y) = AMI(A;x) + AMI(A;y). Every variable is uniformly discretized to have 20 levels at most. Finally we measure averaged mutual information between an attribute and the set of landmarks as

$$\frac{1}{N \times L} \sum_{n \in N} \sum_{x_n \in X_n, y_n \in Y_n} AMI(A_n; x_n) + AMI(A_n; y_n)$$

where N and L indicate the number of samples and landmarks, X_n and Y_n indicate the set of x and y landmark coordinates per sample n. In Table S2 we observe that hand gesture labels and head pose regression are among the most effective attributes for our method. There is little mutual information between wearing glasses and landmarks, indicating lack of usefulness of this attribute for our semi-supervised setting.

Table S2: Mutual Information between all landmarks and each attribute

Dataset		MultiPIE			HGR1
Attribute	Random	Emotion	Camera	Identity	Gesture Label
AMI, mean	.000	.098	.229	.049	.559
AMI, max	.006	.229	.493	.088	.669

Dataset		AFLW	MTFL	
Attribute	Random	Pose Regression	Glasses	Pose Classification
AMI, mean	.000	.536	.002	.069
AMI, max	.006	.576	.003	.222

The attributes that are mostly useful yield a high accuracy, or low error, if we just train a neural network that takes only ground truth landmarks as input and predicts the attribute. This indicates that by relying only on landmarks we can get high accuracy for those attributes. In Table S3 we compare the attribute prediction accuracy from the proposed Seq-MT model with a case when we do such prediction from GT landmarks. Prediction from GT landmarks always outperforms the one of Seq-MT. This indicates that in our semi-supervised setting, where we have few labelled landmarks, by improving the predicted locations of landmarks, both attribute and landmarks error would reduce.

Table S3: Attribute classification accuracy (MultiPIE, HGR1)—higher is better—or prediction error (AFLW)—lower is better—from GT & estimated landmarks.

	MultiPIE		HGR1	AFLW
Attribute	Camera	Emotion	Label	Pose Error
From GT Landmarks	99.54 ↑	88.21 ↑	91.7 ↑	4.98 ↓
Best Seq-MT Attr. Predict.	98.96	86.48	79.1	5.10

S.5.3. Comparison of softmax and soft-argmax

Heatmap-MT(L) and Seq-MT(L) have the same architectures but use different loss functions (softmax vs. soft-argmax). RCN(L) and RCN+(L) also only differ in their loss function. When comparing these models in Tables 1, 2, 3, 5, and 6 soft-argmax outperforms soft-max. To further examine these two losses we replace soft-max with soft-argmax in Heatmap-MT and show the results in Table S4. Comparing the results in Table S4 with Tables 2 and 3, we observe improved performance of landmark localization using soft-armgax. In soft-max the model cannot be more ac-

curate than the number of elements in the grid, since softmax does a classification over the pixels. However, in soft-argmax the model can regress to any real number and hence can get more accurate results. We believe this is the reason behind its better performance.

Table S4: Results on Heatmap-MT (L+A) comparing softmax with soft-argmax.

Dataset		5%	10%	20%	50%	100%
Multi-PIE	softmax	11.03	9.03	8.15	7.11	6.65
	soft-argmax	**8.00**	**7.06**	**6.29**	**5.49**	**5.14**
HGR1	softmax	64.8	54.9	43.2	30.5	26.7
	soft-argmax	**56.88**	**42.79**	**33.07**	**22.5**	**18.8**

S.5.4. Supplementary results on Multi-PIE dataset

Although the focus of this paper is on improving landmark localization, in order to observe the impact of each multi-tasking approach on the attribute classification accuracy, we report the classification results on emotion in Table S5 and on camera in Table S6. Results show that the classification accuracy improves by providing more labeled landmarks, despite having the number of *(image, class label)* pairs unchanged. It indicates that improving landmark localization can directly impact the classification accuracy. Landmarks are especially more helpful in emotion classification. On camera classification, the improvement is small and all models are getting high accuracy. Another observation is that Heatmap-MT performs better on classification tasks compared to the other two multi-tasking approaches. We believe this is due to passing more high-level features from the image to the attribute classification network compared to Seq-MT. However, this model is performing worse than Seq-MT on landmark localization. The Seq-MT model benefits from the landmark bottleneck to improve its landmark localization accuracy. In Tables S5 and S6 by adding the ELT cost the classification accuracy improves (in addition to landmarks) indicating the improved performance in landmark localization can enhance classification performance.

Figure S1 provides further localization examples on Multi-PIE dataset.

S.5.5. Supplementary results on hands dataset

In Table S7 we show classification accuracy obtained using different multi-tasking techniques. Similar to the Multi-PIE dataset, we observe increased accuracy by providing more labeled landmarks, showing the classification would benefit directly from landmarks. Also similar to Multi-PIE, we observe better classification accuracy with Heatmap-MT. Comparing Seq-MT models, we observe improved classification accuracy by using the ELT cost. It demonstrates the impact of this component on both landmark localization and classification accuracy.

Table S5: Emotion classification error on Multi-PIE test set. In percent; higher is better.

	Percentage of Images with Labeled Landmarks				
Model	5%	10%	20%	50%	100%
Comm-MT (L+A)	74.67	79.90	83.76	86.37	86.83
Heatmap-MT (L+A)	**85.14**	**87.50**	**86.93**	**88.16**	**87.29**
Seq-MT (L+A)	78.78	82.62	84.69	84.03	84.86
Seq-MT (L+A+ELT)	82.90	84.57	84.85	86.48	

Table S6: Camera classification error on Multi-PIE test set. In percent; higher is better.

	Percentage of Images with Labeled Landmarks				
Model	5%	10%	20%	50%	100%
Comm-MT (L+A)	96.98	97.53	98.30	98.63	98.80
Heatmap-MT (L+A)	**98.46**	**98.99**	**98.99**	**98.98**	**98.98**
Seq-MT (L+A)	97.97	98.31	98.50	98.96	98.92
eq-MT (L+A+ELT)	98.41	98.53	98.47	98.43	

Table S7: Classification error on hands test set. In percent; higher is better.

	Percentage of Images with Labeled Landmarks				
Model	5%	10%	20%	50%	100%
Comm-MT (L+A)	60.86	69.64	69.20	76.03	73.42
Heatmap-MT (L+A)	**83.74**	**87.86**	**87.55**	**90.29**	**89.27**
Seq-MT (L+A)	69.08	70.14	72.26	77.07	75.92
Seq-MT (L+A+ELT)	74.64	75.01	73.90	79.10	

Figure S2 provides further landmark localization examples on hands dataset.

S.5.6. Supplementary results on 300W dataset

In Figure S3 we show the architecture of *RCN* $^+$ used for 300W and AFLW datasets. In Figure S4 we illustrate further samples from 300W dataset. The samples show the improved accuracy obtained in both *Seq-MT* and *RCN* $^+$ by using the ELT loss.

S.5.7. Supplementary results on AFLW dataset

In Table S8 we show pose estimation error using different percentage of labelled data for RCN$^+$ (L+ELT+A) model and compare the results to a model trained to estimate pose from GT landmarks. All models get close results compared to GT model indicating RCN$^+$ (L+ELT+A) can do a reliable pose estimation using a small set of labelled landmarks.

Figure S5 shows some samples on AFLW test set.

S.5.8. Architecture details

The architecture details of Seq-MT model on different datasets can be seen in Tables S11, S12 and S13. Architecture details of Comm-MT and Heatmap-MT for Blocks dataset are shown in Tables S9 and S10. For other dataset,

Figure S1: Extra examples of our model predictions on Multi-PIE [10] test set. We observe close predictions by 1) and 2) indicating the effectiveness of our proposed ELT cost even with only a small amount of labeled landmarks. Comparison between 3) and 4) shows the improvement obtained with both the ELT loss and the sequential multitasking architecture when using a small percentage of labeled landmarks. Note that the model trained with ELT loss preserves better the joint distribution over the landmarks even with a small number of labeled landmarks. The last two examples show examples with high errors. Best viewed in color with zoom.

Table S8: Pose degree estimation error on AFLW test set, as average of yaw, pitch, roll values. lower is better.

	Percentage of Images with Labeled Landmarks		
Model	1%	5%	100%
RCN$^+$(L+ELT+A)	5.05	5.01	5.12
GT			4.98

Table S9: Architecture details for Comm-MT Model on Blocks dataset.

Input = $60 \times 60 \times 1$	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Pool 2×2, stride 2	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Pool 2×2, stride 2	
Conv $1 \times 1 \times 8$, ReLU, stride 1, SAME	
Conv $1 \times 1 \times 8$, ReLU, stride 1, SAME	
FC $\#units = 256$, ReLU, dropout-prob=.25	
FC $\#units = 256$, ReLU, dropout-prob=.25	
Classification branch	Landmark localization branch
FC $\#units = 15$, Linear softmax(dim=15)	FC $\#units = 10$, Linear

Table S10: Architecture details for Heatmap-MT Model on Blocks datasets.

Input = $60 \times 60 \times 1$	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	
Conv $1 \times 1 \times 8$, ReLU, stride 1, SAME	
Conv $1 \times 1 \times 5$, ReLU, stride 1, SAME	
classification branch	landmark localization branch
Pool 2×2, stride 2	—
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	—
Pool 2×2, stride 2	—
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	—
Pool 2×2, stride 2	—
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	—
Pool 2×2, stride 2	—
Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME	—
FC $\#units = 256$, ReLU, dropout-prob=.25	—
FC $\#units = 256$, ReLU, dropout-prob=.25	—
FC $\#units = 15$, Linear	—
softmax(dim=15)	softmax(dim=60×60)

the kernel size and the number of feature maps for conv layers and the number of units for FC layers change similar to Seq-MT model on those datasets.

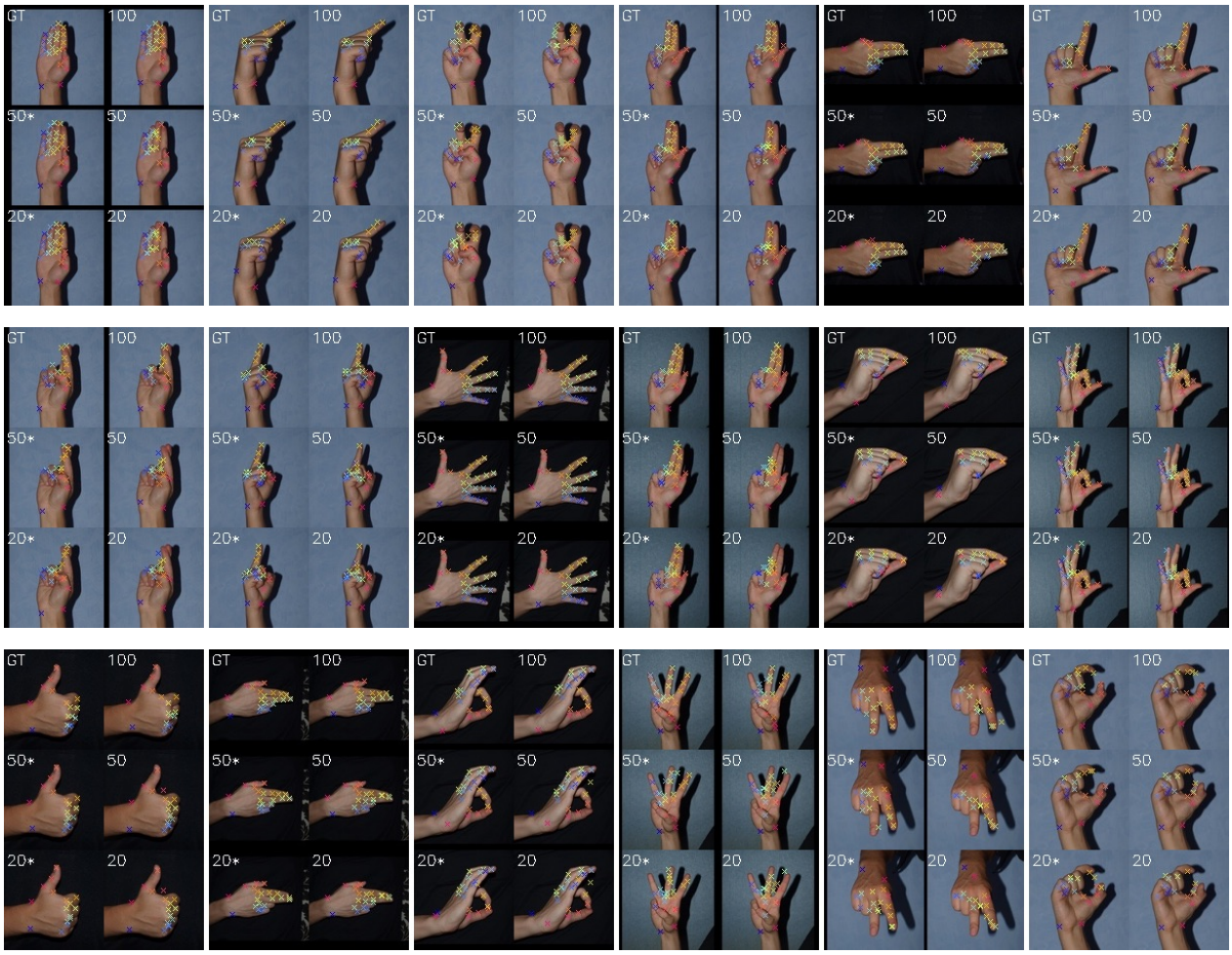

Figure S2: Extra examples of our model predictions on the HGR1 [16, 23] test set. GT represents ground-trust annotations, while numbers 100, 50, and 20 indicate the percentage of the training set with labeled landmarks. Results are computed with Seq-MT (L+ELT+A) model (denoted *) and Seq-MT (L). Examples illustrate improvement of the landmark prediction by using the class label and the ELT cost in addition to the labeled landmarks. The last three examples on the bottom row show examples with high errors. Best viewed in color with zoom.

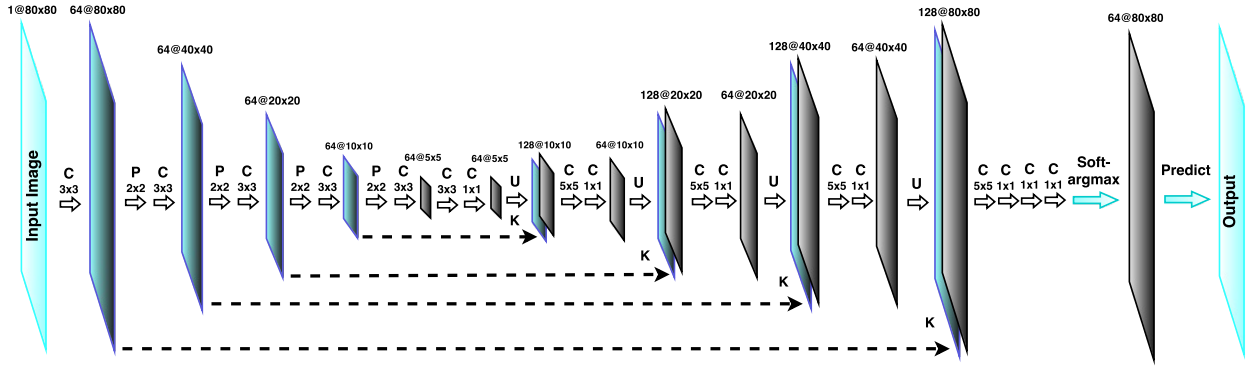

Figure S3: The ReCombinator Networks (RCN) [13] architecture used for experiments on 300W dataset. P indicates a pooling layer. All pooling layers have stride of 2. C indicates a convolutional layer. The number written below C indicates the convolution kernel size. All convolutions have stride of 1. U indicates an upsampling layer, where each feature map is upsampled to the next (bigger) feature map resolution. K indicates concatenation, where the upsampled features are concatenated with features of the same resolution before a pooling is applied to them. The dashed arrows indicate the feature maps are carried forward for concatenation. The solid arrows following each other, e.g. P, C, indicate the order of independent operations that are applied. The number written above feature maps in $n@w \times h$ format indicate number of feature maps n and the width w and height h of the feature maps. On AFLW, we use 70 feature maps per layer (instead of 64) and we get two levels coarser to get to 1×1 resolution (instead of 5×5). On both datasets we shoud $\beta = 100$ for soft-argmax layer.

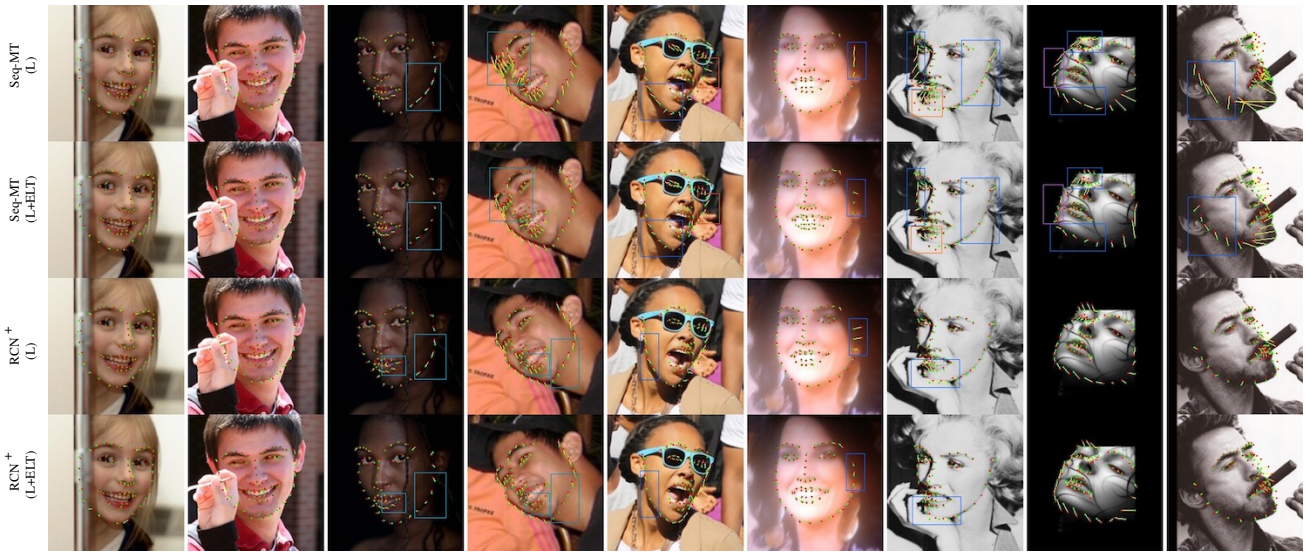

Figure S4: Extra examples of our model predictions on 300W [27] test-set. The first two columns depict examples where all models get accurate predictions, The next 5 columns illustrate the improved accuracy obtained by using ELT loss in two different architectures (Seq-MT and RCN). The last two columns show difficult examples where error is high. The rectangles indicate the regions that landmarks are mostly affected. The green and red dots show ground truth (GT) and model predictions (MP), respectively. The yellow lines show the error by connecting GT and MP. Note that the ELT loss improves predictions in both architectures. Best viewed in color with zoom.

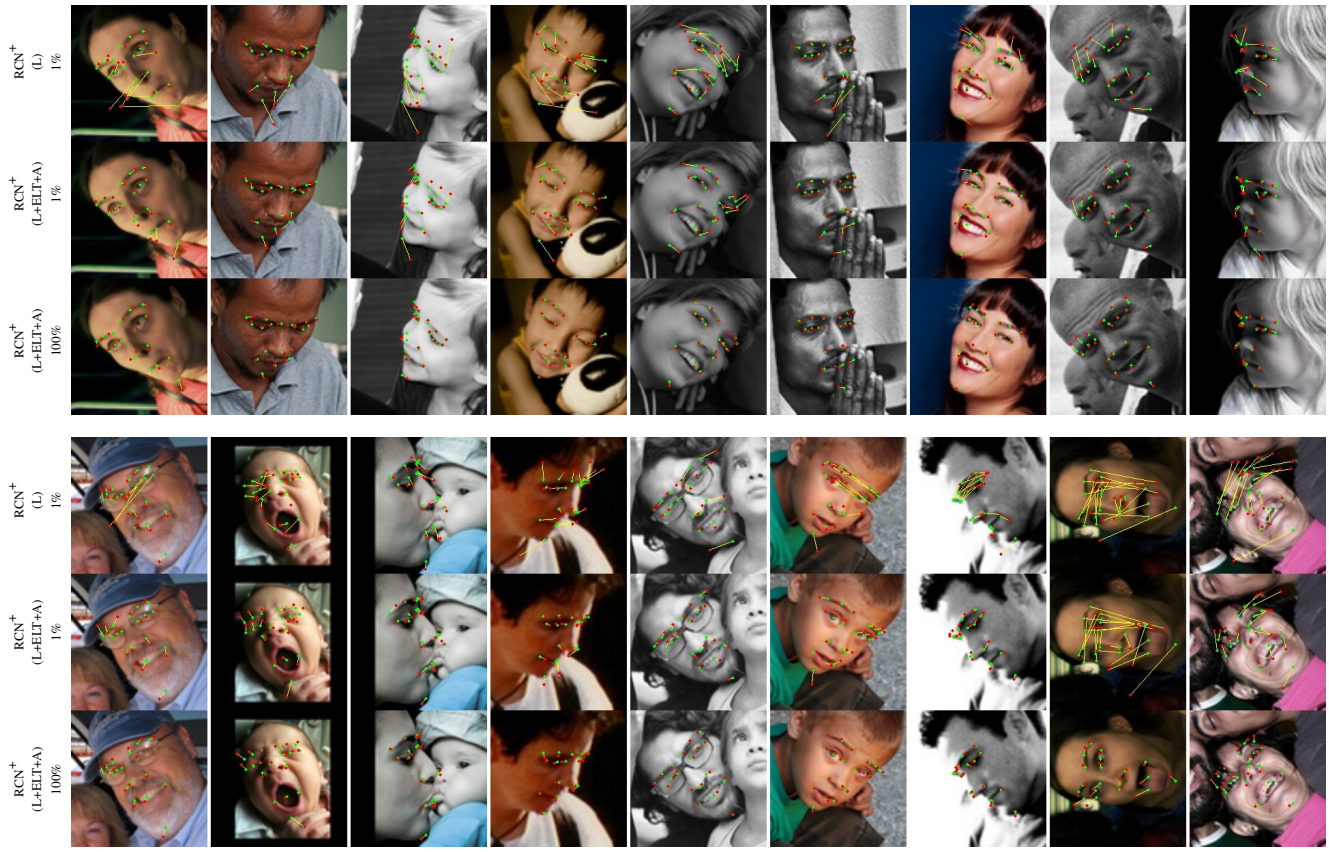

Figure S5: Extra examples of our model predictions on the AFLW test set. Comparing the first and second rows shows the improvement obtained by using ELT+A with only 1% of labelled landmarks. Note the model trained using ELT+A preserves better the distribution over the landmarks. The last two columns in the bottom row show samples with high error on small percentage of labelled landmaks, which is due to extreme rotation. The bottom row shows the prediction using L+ELT+A on the entire set of labelled landmarks, which gets the best results. The green and red dots show ground truth (GT) and model predictions (MP), respectively. The yellow lines show the error by connecting GT and MP. Best viewed in color with zoom.

Table S11: Architecture details of Seq-MT model used for Shapes and Blocks datasets. Each conv layer has three values as $w \times h \times n$ indicating width (w), height (h) of kernel and the number of feature maps (n) of the convolutional layer. SAME indicates the input map is padded with zeros such that input and output maps have the same resolution.

Shapes Dataset	**Blocks Dataset**
Model HP: $\lambda = 0, \alpha = 0, \gamma = 0, \beta = 1$, ADAM	Model HP: $\lambda = 1, \alpha = 1, \beta = 1$, ADAM
Landmark Localization Network	Landmark Localization Network
Input = $60 \times 60 \times 1$ Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $7 \times 7 \times 16$, ReLU, stride 1, SAME Conv $1 \times 1 \times 16$, ReLU, stride 1, SAME Conv $1 \times 1 \times 2$, ReLU, stride 1, SAME soft-argmax(num_channels=2)	Input = $60 \times 60 \times 1$ Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $9 \times 9 \times 8$, ReLU, stride 1, SAME Conv $1 \times 1 \times 8$, ReLU, stride 1, SAME Conv $1 \times 1 \times 5$, ReLU, stride 1, SAME soft-argmax(num_channels=5)
Classification Network	Classification Network
FC $\#units = 40$, ReLU FC $\#units = 2$, Linear softmax(dim=2)	FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 15$, Linear softmax(dim=15)

Table S12: Architecture details of Seq-MT model used for Hands and Multi-PIE datasets.

Hands Dataset	**Multi-PIE Dataset**	
Model HP: $\lambda = 0.5, \alpha = 0.3, \gamma = 10^{-5}, \beta = 0.001$, ADAM	Model HP: $\lambda = 2, \alpha = 0.3, \gamma = 10^{-5}, \beta = 0.001$, ADAM	
Preprocessing: scale and translation [-10%, 10%] of face bounding box, rotation [-20, 20] applied randomly to every epoch.		
Landmark Localization Network	Landmark Localization Network	
Input = $64 \times 64 \times 1$ Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 25$, ReLU, stride 1, SAME soft-argmax(num_channels=25)	Input = $64 \times 64 \times 1$ Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 64$, ReLU, stride 1, SAME Conv $9 \times 9 \times 68$, ReLU, stride 1, SAME soft-argmax(num_channels=68)	
Classification Network	Emotion Classification Branch	Camera Classification Branch
FC $\#units = 256$, ReLU, dropout-prob=.5 FC $\#units = 256$, ReLU, dropout-prob=.5 FC $\#units = 27$, Linear softmax(dim=27)	FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 6$, Linear softmax(dim=6)	FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 256$, ReLU, dropout-prob=.25 FC $\#units = 5$, Linear softmax(dim=5)

Table S13: Architecture details of Seq-MT model used for 300W datasets.

300W Dataset
Model HP: $\lambda = 2.0, \alpha = 2.0, \gamma = 10^{-5}, \beta = 0.01$, ADAM
Preprocessing: scale and translation [-10%, 10%] of face bounding box, rotation [-30, 30] applied randomly to every epoch.
Landmark Localization Network
Input = $64 \times 64 \times 1$ Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 32$, ReLU, stride 1, SAME Conv $9 \times 9 \times 68$, ReLU, stride 1, SAME soft-argmax(num_channels=68)