Wang, Tzu-Jui Julius; Laaksonen, Jorma; Liao, Yi-Ping; Wu, Bo-Zong; Shen, Shih-Yun

A Multi-Task Bayesian Deep Neural Net for Detecting Life-Threatening Infant Incidents From Head Images

Published in:
2019 IEEE International Conference on Image Processing, ICIP 2019 - Proceedings

DOI:
10.1109/ICIP.2019.8803332

Published: 01/01/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Wang, T.-J. J., Laaksonen, J., Liao, Y.-P., Wu, B.-Z., & Shen, S.-Y. (2019). A Multi-Task Bayesian Deep Neural Net for Detecting Life-Threatening Infant Incidents From Head Images. In *2019 IEEE International Conference on Image Processing, ICIP 2019 - Proceedings* (pp. 3006-3010). Article 8803332 IEEE. https://doi.org/10.1109/ICIP.2019.8803332

A MULTI-TASK BAYESIAN DEEP NEURAL NET FOR DETECTING LIFE-THREATENING INFANT INCIDENTS FROM HEAD IMAGES

Tzu-Jui (Julius) Wang[*], *Jorma Laaksonen*[*] *Yi-Ping Liao* [†] *Bo-Zong Wu* [‡], *Shih-Yun Shen* [‡]

[*]Aalto University, Department of Computer Science, Finland

[‡]Yun yun AI Baby camera Co,. Ltd., Taiwan (R.O.C)

ABSTRACT

The notorious incident of sudden infant death syndrome (SIDS) can easily happen to a newborn due to many environmental factors. To prevent such tragic incidents from happening, we propose a multi-task deep learning framework that detects different facial traits and two life-threatening indicators, i.e. which facial parts are occluded or covered, by analyzing the infant head image. Furthermore, we extend and adapt the recently developed models that capture data-dependent uncertainty from noisy observations for our application. The experimental results show significant improvements on YunInfants dataset across most of the tasks over the models that simply adopt the regular cross-entropy losses without addressing the effect of the underlying uncertainties.

***Index Terms*—** Bayesian deep neural net, occlusion detection, cover detection, neonate safety

1. INTRODUCTION

Facial analysis has always been one of the core problems in the computer vision field. In particular, problems such as face detection [1], facial landmark detection [2], and pose estimation [3] have been progressing at an unprecedented pace thanks to the advancements in deep learning approaches. That being said, most efforts have been poured into analyzing the subjects who are either children or adults [4, 5], but much less have been focused on the infant group, which is considered vulnerable and requires being attended. This may have stagnated the development in computer vision applications centered around the infants. For instance, in surveillance applications, while many more systems are rather general purpose, they are not crafted for monitoring neonatal safety. Having such a safety system is crucial since the notorious incident of sudden infant death syndrome (SIDS) can happen quite easily to a newborn under several risky conditions [6].

In this work, as one crucial step towards protecting a newborn against SIDS, we propose a multi-task Bayesian deep neural architecture consisting of four different sub-tasks of recognizing: a) whether the eyes, nose, or mouth, are *occluded*, b) whether they are *covered* by arbitrary objects, c) whether the eyes are open, and d) the locations of the five facial landmarks. A facial part is defined to be *occluded* when not visible. A facial part is defined to be *covered* when not visible and covered by external objects, such as a pillow. Having both occlusion and cover detection tasks in the same learning framework makes it possible to analyze whether the occlusion is caused by sleeping position or the surrounding objects. Being not visually apparent, analyzing infant facial traits can be challenging as shown in Fig. 1. We would like to develop a system that can assist any monitoring system in telling if an infant is in an improper sleeping position, causing her face being covered by the bedding. Existing works inferring occlusion on head images, such as [7, 8, 9], do not focus on analyzing infant images. On the contrary, we introduce a novel dataset named YunInfants and propose a method focusing on the infant group, aiming to handle any head pose and imaging under varying lighting conditions.

To be applied in a system that considers safety issues, it

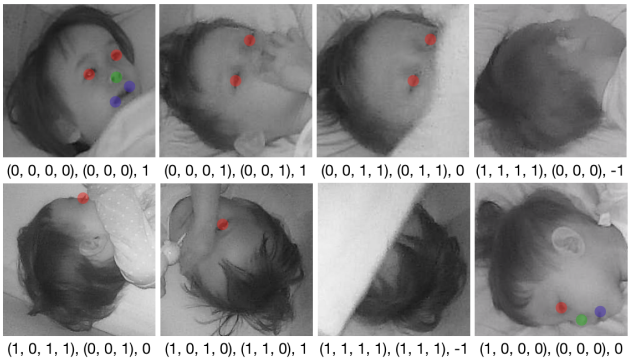

Fig. 1. Instances of annotations for the four tasks on infant head images on YunInfants. Labels for each image contain three parts. 1/0 in the first 4-element vector (and the second 3-element vector) encode whether an eye, eyes, nose and mouth are occluded (covered)/not occluded (not covered). 1/0 in the third scalar element says the eyes are open / not open.

[*]Email: {tzu-jui.wang, Jorma.Laaksonen}@aalto.fi
[†]Equal contribution with Julius Wang. Email: yiping.ncu@gmail.com
[‡]Email: {barry, steven}@yunyun.cloud

is suggested that the network should estimate uncertainty of its predictions [10]. Two types of uncertainties can be considered in a learning task, i.e. *aleatoric* and *epistemic* uncertainties [10]. We emphasize the former one as it comes from the inherent noise in the observation, such as the noise or the sensing capability of the sensor, and can lead to ambiguous predictions. This reflects the scenario in our application because the collected images contain noises due to: 1) the camera switching to the night-vision mode, where its sensing ability is limited and often noisy, and 2) the motion blur caused by the sudden movement of the baby.

To summarize the contributions in this paper: a) We propose a multi-task Bayesian deep learning framework that learns four detection tasks accounting for neonatal safety. This is a critical and, to our knowledge, novel application in computer vision. b) We extend the loss function for classification defined in [10] to multi-label and multi-task problems. Empirically we show that learning with this extended loss function accounting for data-dependent uncertainty brings significant improvement in nearly all tasks over the baselines on the collected YunInfants dataset. The following sections are structured as follows. In Sec. 2, we introduce the proposed network architecture and formalize the objectives of different learning tasks in the network. In Sec. 3, we present the experimental results, which are then followed by conclusions in Sec. 4.

2. PROPOSED METHOD

We propose a Bayesian multi-task neural network for four different detection tasks, including a) facial occlusion, b) facial cover, c) eye openness, and d) five-point facial landmark detection.

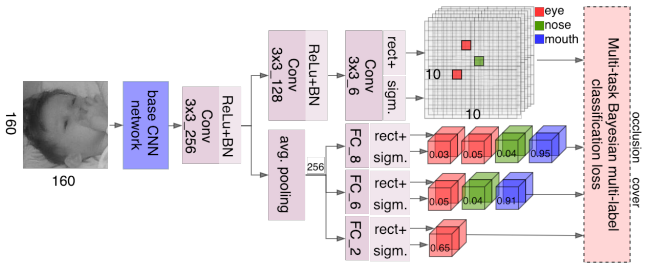

Fig. 2. The proposed multi-task network architecture for occlusion, cover, eye openness, and landmark detection tasks. `Convkxk_n` represents a convolution layer with kernel size `k×k`, `n` output channels, and stride being 1. `FC_n` represents a fully-connected layer with `n` outputs. `sigm.` and `rect+` indicates sigmoid activation function and the rectifier that enforce positiveness, respectively. `BN` indicates batch normalization. The network, in each task branch, estimates not only the posterior mean of the class distribution, but also the uncertainty of the estimates.

2.1. Network Architecture and Objectives

2.1.1. Labels

Occlusion, Cover, and Eye Openness Detection: The labels defined for occlusion, cover, and eye openness detection tasks are denoted as $\mathbf{y}_{occ} \in \{0,1\}^4$, $\mathbf{y}_{cov} \in \{0,1\}^3$, and $y_{eye} \in \{0,1\}$ respectively. 1's (0's) in these labels represent if a facial part is occluded (not occluded), covered (not covered), or open (closed), respectively for $\mathbf{y}_{occ}$, $\mathbf{y}_{cov}$, and y_{eye}. The conditions of the four facial parts (an eye, both eyes, nose, mouth) are enclosed in $\mathbf{y}_{occ}$, while those of three facial parts (eyes, nose, mouth) are captured in $\mathbf{y}_{cov}$. y_{eye} captures the eye openness. One can refer to Fig. 1 for labeling examples. Note that if a human annotator cannot tell the condition, then it is labeled as -1. We follow the same landmark annotation protocol used in [5] except that we do not annotate the occluded facial part.

Five-Point Facial Landmark Detection: Most of the works on landmark detection assume that every landmark is visible regardless of the pose and occlusion [5]. However, it is not a suitable setting in the application addressed in this paper since one can easily find a head image visible with only few or no landmarks (as in some examples in Fig. 1). Hence the task here is not to estimate the coordinates of the landmarks, but to detect if those five landmarks are present in the grid space $Y_{lm} \in \{0,1\}^{W_{lm} \times H_{lm} \times 3}$, where the three $W_{lm} \times H_{lm}$ landmark maps respectively correspond to the visibility of eyes, nose and mouth at a lower resolution of the input image.

2.1.2. Inputs and Outputs

Fig. 2 depicts the overview of the proposed framework. The network takes a head image as input $X \in \mathbb{R}^{w_i \times h_i \times c_i}$, which can come from any head detector. The outputs of the network are collected from four task branches, each of which predicts the probability of the class labels and the their observation noises. Concretely, the landmark, occlusion (occ), cover (cov), and eye openness (eye) branches (subscripted by lm, occ, cov, and eye, respectively) predict ($\mathbf{y}'_{occ} \in [0,1]^4$, $\boldsymbol{\sigma}_{occ} \in \mathbb{R}^4_{\geq 0}$), ($\mathbf{y}'_{cov} \in [0,1]^3$, $\boldsymbol{\sigma}_{cov} \in \mathbb{R}^3_{\geq 0}$), ($y'_{eye} \in [0,1]$, $\sigma_{eye} \in \mathbb{R}_{\geq 0}$), and ($Y'_{lm} \in [0,1]^{W_{lm} \times H_{lm} \times 3}$, $S_{lm} \in \mathbb{R}^{W_{lm} \times H_{lm} \times 3}_{\geq 0}$), respectively.

2.1.3. Network Architecture

The proposed network architecture is shown in Fig. 2. The input image X is fed into a CNN base network $\mathcal{F}_{base} : \mathbb{R}^{160 \times 160 \times 3} \rightarrow \mathbb{R}^{w_b \times h_b \times c_b}$. The output of the base network, $\mathcal{F}_{base}(X)$, flows through the operations, C_{shar}, consisting of shared convolution layer followed by ReLu and batch normalization, yielding $Z = C_{shar}(\mathcal{F}_{base}(X))$, which is the input to the succeeding detection tasks, where the computations are carried out through functions $\mathcal{F}_{lm}$, $\mathcal{F}_{occ}$, $\mathcal{F}_{cover}$, and $\mathcal{F}_{eye}$, respectively, defined as follows (* denotes either

occ, cov, eye):

$$\mathcal{F}_{lm} = C_{lm,2}(C_{lm,1}(Z)), \quad (1)$$

$$\mathcal{F}_{*} = FC_{*}(O(Z)), \quad (2)$$

where $O(\cdot)$ is the global average pooling used in [11]. $\mathcal{F}_{lm}$, the predictions in the landmark branch is defined by convolutional layers $C_{lm,1}$ and $C_{lm,2}$, in which the latter outputs Y'_{lm} and S_{lm}. FC_{occ}, FC_{cov}, and FC_{eye} are fully-connected layers with eight, six, and two output neurons, respectively. Among the outputs of each FC_{*}, the first half of neurons are the logit values followed by the sigmoid activation functions for estimating the class probabilities. The second half of neurons are the values which are then activated by softplus function [12] that enforces positiveness for estimating the standard deviations of the class probabilities. In the following subsection, we formulate the losses for the different tasks.

2.1.4. Losses with Aleatoric Uncertainty

The model is learned by minimizing L, the sum of the weighted losses, i.e.

$$L = \alpha_{occ}L_{occ} + \alpha_{cov}L_{cov} + \alpha_{eye}L_{eye} + \alpha_{lm}L_{lm}, \quad (3)$$

where each loss term for each task is introduced as follows.

Multi-Label Losses: We formulate the multi-label classification problem as the composition of $c_{occ} = 4$, $c_{cov} = 3$, and $c_{eye} = 1$ binary classification problem(s) for occlusion, cover, eye openness detection tasks, respectively. c_{*} is the number of classes in the corresponding detection task. This allows one to easily extend the classification loss with heteroscedastic aleatoric uncertainty proposed in [10] to multi-label classification tasks. Concretely, given a target vector $\mathbf{y}_{*} = [y_{*,i}]_{i=1,...,c_{*}}$, we apply one-hot transform on $\mathbf{y}_{*}$ to obtain $Y_{*} = [\mathbf{y}^{T}_{*,i}]_{i=1,...,c_{*}} \in \mathbb{R}^{c_{*}\times 2}$, where $\mathbf{y}^{T}_{*,i} = [1, 0]$ if $y_{*,i} = 0$, otherwise, $\mathbf{y}^{T}_{*,i} = [0, 1]$. Likewise, we define a vector of predicted logit values as $\mathbf{f}_{*} = [f_{*,i}]_{i=1,...,c_{*}}$, where

$$\mathbf{y}'_{*} = \sigma(\mathbf{f}_{*}) = [\sigma(f_{*,i})]_{i=1,...,c_{*}}, \quad (4)$$

$\sigma(\cdot)$ is the sigmoid function, and $\mathbf{y}'_{*}$ are defined in Sec. 2.1.2. Next, one can transform $f_{*,i}$ into a matrix, $F_{*} = [\mathbf{f}^{T}_{*,i}]_{i=1,...,c_{*}}$, where $\mathbf{f}^{T}_{*,i} = [-f_{*,i}, f_{*,i}]$.

To compute the loss with aleatoric uncertainty, one generates N noisy logit outputs $\hat{\mathbf{y}}^{t}_{*}$ from $\mathbf{f}_{*}$ for $t = 1, ..., N$, where

$$\hat{\mathbf{y}}^{t}_{*} = \mathbf{f}_{*} + \boldsymbol{\sigma}_{*}\epsilon^{t},\ \epsilon^{t} \sim \mathcal{N}(0, 1),\ \hat{\mathbf{y}}^{t}_{*} = [\hat{y}^{t}_{*,i}]_{i=1,...,c_{*}}. \quad (5)$$

From Eq. (5), one can apply the transformation applied to $f_{*,i}$ as described above to obtain $\hat{Y}^{t}_{*} = [\hat{\mathbf{y}}^{tT}_{*,i}]_{i=1,...,c_{*}}$, $\hat{\mathbf{y}}^{tT}_{*,i} = [-\hat{y}^{t}_{*,i}, \hat{y}^{t}_{*,i}]$. Then one can compute the multi-label loss per sample with aleatoric uncertainty using

$$L_{*} = -\log \frac{1}{N}\sum_{t=1}^{N}\frac{1}{c_{*}}\sum_{i=1}^{c_{*}}\exp\big(\mathbf{y}^{T}_{*,i}\cdot\hat{\mathbf{y}}^{tT}_{*,i} - \log(\exp(-\hat{y}^{t}_{*,i}) + \exp(\hat{y}^{t}_{*,i}))\big). \quad (6)$$

Auxiliary Loss on Facial Landmark: Adding an auxiliary loss L_{lm} defined over facial landmarks allows the model to learn to see the facial features. In addition, we argue that having the landmark detection results available is useful for model diagnosis as shown later in the experiments. The landmark loss calculation is exactly the same as the other tasks except that one has to flatten Y'_{lm} and Y_{lm} before feeding them through the computations in Eqs. (5) and (6).

3. EXPERIMENTS AND RESULTS

3.1. Dataset

To the best of our knowledge, there does not exist any publicly accessible dataset similar to YunInfants that we collected from over a hundred users under their consent. Table 1 shows the data statistics. YunInfants, containing both day and night-vision head images, is recorded in home environments with the camera pointing to the crib. The images greatly vary in head poses, lighting conditions, and sometimes suffer from motion blur. Most of the subjects (i.e. infants) are Asian.

3.2. Network Parameters

Inputs: All input images are resized to a fixed size of $160 \times 160 \times 3$. The width and height of the ground-truth landmark maps are set to be $(W_{lm}, H_{lm}) = (10, 10)$.

Hyperparameters: The base CNN network we adopt is MobileNetV2 [13] for computational efficiency. We extract features from the 14th convolution layer, reducing the size to $\frac{1}{16}$ of the original image in both width and height. This gives us the feature maps of size $10 \times 10 \times 96$. The weights $(\alpha_{occ}, \alpha_{cov}, \alpha_{eye}, \alpha_{lm})$ used to define the total loss in Eq. (3) are set to be (10, 1, 1, 1). We set $T = 25$ used in Eq. (6) throughout all the experiments that require computing the losses that capture data-dependent uncertainty. All the models are trained with 130k batches (50 images / batch), and the best model of each method used in the test phase is picked according to the average validation accuracy (in terms of the F1-score) at 90k, 100k, ..., and 130k batches.

3.3. Experiments

Here we study the effectiveness of the proposed method by comparing the models in the multi-task learning setting against those in the single-task setting, and the models considering uncertainty against those not. Though other metrics measuring the multi-label task are available, we report the F1-scores (i.e. the harmonic mean of precision and recall) on all the classes in each task in Table 2 to allow analyzing individual classification results. F1-score is adopted here to address the imbalanced class distributions as shown in Table 1. Comparing single-task models, we find that they are on a par with each other across the three tasks, no matter if they

Table 1. Statistics of the training, validation, and test data.

# of images in train/val/test splits: 12850/3211/3655. The numbers in each cell show the statistics of these splits.										
% occl.	eye	eyes	nose	mouth	% cover.	eye	nose	mouth	% eye open.	eye
	37/38/38	21/21/15	38/38/32	58/58/54		12/13/13	21/21/24	34/35/39		22/19/20

Table 2. Comparison on F1-scores obtained from different models on three tasks: occlusion ("o"), cover ("c"), and eye openness ("e") detection tasks. The models vary with: 1) whether it is of single-task (presented in the first two rows), where it is optimized solely on selected tasks, or multi-task (presented in the last four rows), where it is optimized on all the tasks, 2) whether Bayesian auxiliary landmark loss ("baux") is employed, and 3) whether the losses are those consider uncertainty ("bayes") or regular cross-entropy losses ("xent") used for measuring classification errors.

	occlusion					**cover**				**eye**	**all tasks**
F1-score (%)	eye	eyes	nose	mouth	avg.	eye	nose	mouth	avg.	**openness**	avg.
{o,c,e}+baux	80.71	75.64	83.24	89.97	82.39	69.52	78.01	81.11	76.21	82.56	80.39
{o,c,e}+baux+bayes	80.01	72.06	84.09	89.63	81.45	70.78	77.12	80.91	76.27	83.20	80.31
o+c+e+bayes	79.96	76.76	83.93	90.14	82.70	72.71	78.72	81.50	77.64	83.29	81.21
o+c+e+baux	80.26	75.02	83.56	90.00	82.23	71.17	79.00	81.76	77.31	**84.51**	81.35
o+c+e+baux+bayes	**82.19**	**77.16**	**85.64**	**91.09**	**84.02**	**74.82**	**79.57**	**83.11**	**79.17**	83.92	**82.37**

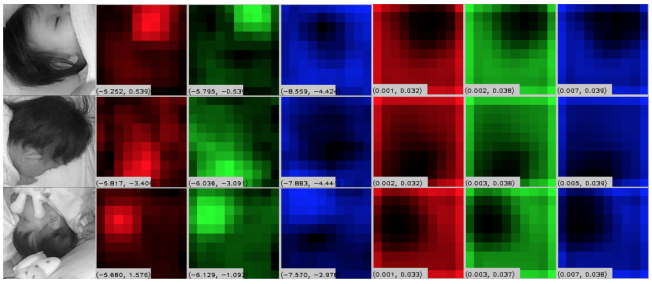

Fig. 3. Qualitative results on correctly classified examples. Each row shows, firstly, the input image, secondly, from the second to fourth columns, the landmark detection results for eyes, nose, and mouth, respectively, and thirdly, from the fifth to seventh columns, the predicted uncertainty maps for aforementioned three facial parts respectively. The numbers in landmark and uncertainty maps are the minimum and maximum in logit values and predicted standard deviations, respectively. The labels for the image in each row are, from top to bottom, ([0, 0, 1, 1], [0, 1, 1], 0), ([1, 1, 1, 1], [0, 0, 0], 0), and ([1, 0, 1, 1], [0, 1, 1], 0), respectively. One can read these tuples by referring to the labeling format defined in Fig. 1.

model uncertainty or not. One can observe the improvements across tasks in the multi-task models listed. The multi-task model trained with losses capturing uncertainty along with the auxiliary landmark losses stably obtains the highest F1-scores in nearly all tasks except in the eye openness detection task. This demonstrates the effectiveness of the proposed multi-task models. Fig. 3 shows that the proposed model is capable of differentiating whether a facial part is occluded or covered. From the same figure, one can also perceive that the predicted landmark maps still produce higher logit values around facial parts even under occlusion. In addition, smaller uncertainty values clutter around the facial area while larger uncertainty values are found in the background. This can be explained by the fact that comparing to the areas that contain targets, i.e. faces, the background areas are of greater visual variability and the model cannot there be certain with its predictions. In other words, the facial landmark detection and uncertainty maps combined act as a face detector which can provide evident visual cues on which other relevant tasks can hinge to make predictions.

4. CONCLUSIONS

This paper addressed detection problems with emphasis on infant head images, including occlusion, cover, eye openness, and landmark detection. We introduced YunInfants dataset and propose a Bayesian deep learning framework trained with the loss functions that consider uncertainties across different detection tasks while utilizing the auxiliary landmark data to further enhance the detection accuracy in terms of the F1-score. The qualitative results on the collected YunInfants dataset also show that the estimated uncertainties capture meaningful signals from the head images.

This work can be extended in different directions. To name one possibility, one can directly learn from the data how to weigh different losses from multiple tasks as suggested in [14]. In addition, we plan to add more tasks into the proposed framework to fully exploit the benefit of learning the representations shared across different, but relevant visual tasks.

5. ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland project number 313988, computational resources provided by the Aalto Science-IT project and NVIDIA Corporation.

6. REFERENCES

[1] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[2] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3067–3074, 2018.

[3] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[4] Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W Picard, "Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 339–346.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[6] Henry F Krous, J Bruce Beckwith, Roger W Byard, Torleiv O Rognum, Thomas Bajanowski, Tracey Corey, Ernest Cutz, Randy Hanzlick, Thomas G Keens, and Edwin A Mitchell, "Sudden infant death syndrome and unclassified sudden infant deaths: a definitional and diagnostic approach," *Pediatrics*, vol. 114, no. 1, pp. 234–238, 2004.

[7] Ville Viitaniemi, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen, "Detecting hand-head occlusions in sign language video," in *Scandinavian Conference on Image Analysis*. Springer, 2013, pp. 361–372.

[8] Golnaz Ghiasi, Charless C Fowlkes, and C Irvine, "Using segmentation to predict the absence of occluded parts.," in *BMVC*, 2015, pp. 22–1.

[9] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo, "Detecting masked faces in the wild with lle-cnns," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2682–2690.

[10] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[12] Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li, "Improving deep neural networks using softplus units," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–4.

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520.

[14] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.