# EFFICIENT GROUP-BASED COHESION PREDICTION IN IMAGES USING FACIAL DESCRIPTORS

**Ilya Gavrikov**
National Research University Higher School of Economics – Nizhny Novgorod, student
Email: ilsgavrikov@gmail.com

**Andrey V. Savchenko**
Dr. of Sci., Prof., LATNA, National Research University Higher School of Economics – Nizhny Novgorod
Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko

# MOTIVATION

(a) Strongly disagree

(b) Disagree


(c) Agree

(d) Strongly agree

(a) Negative

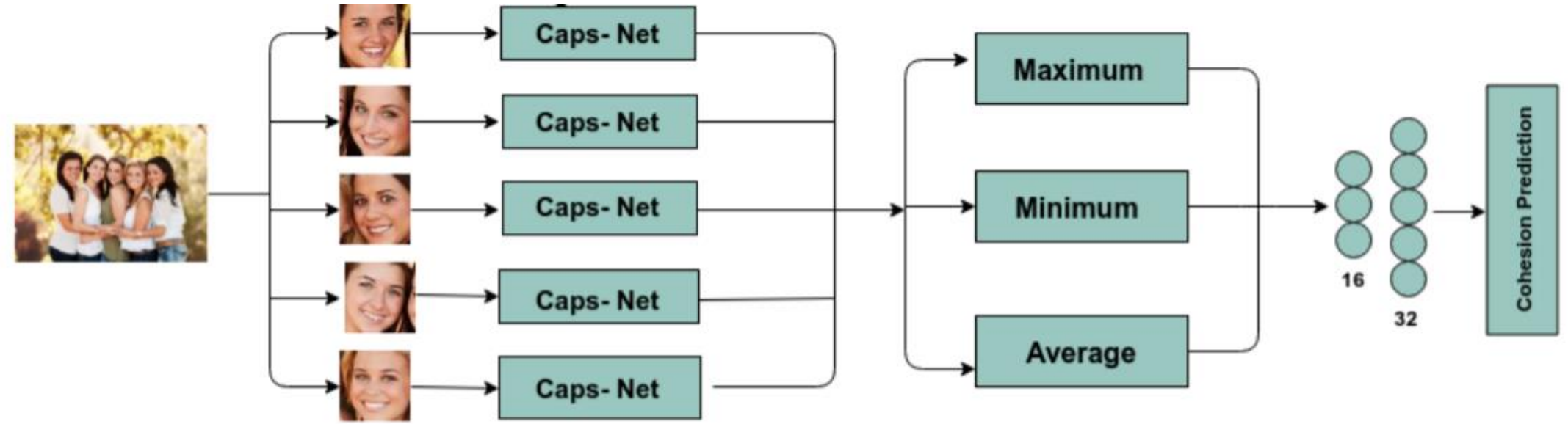
(b) Neutral

(c) Positive

Ghosh, S., Dhall, A., Sebe, N., Gedeon, T.: Predicting group cohesiveness in images. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). pp. 1{8. IEEE (2019)

# REVIEW

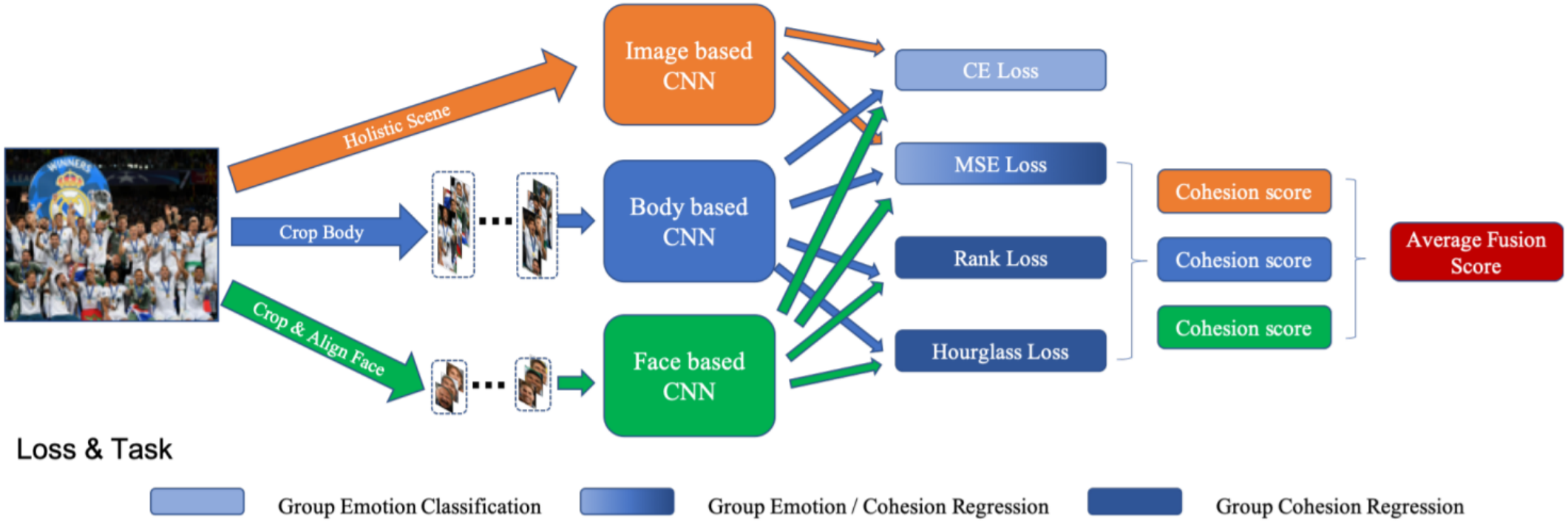
**MSE = 0.84**




Ghosh, S., Dhall, A., Sebe, N., Gedeon, T.: Predicting group cohesiveness in images. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). pp. 1{8. IEEE (2019)

# REVIEW

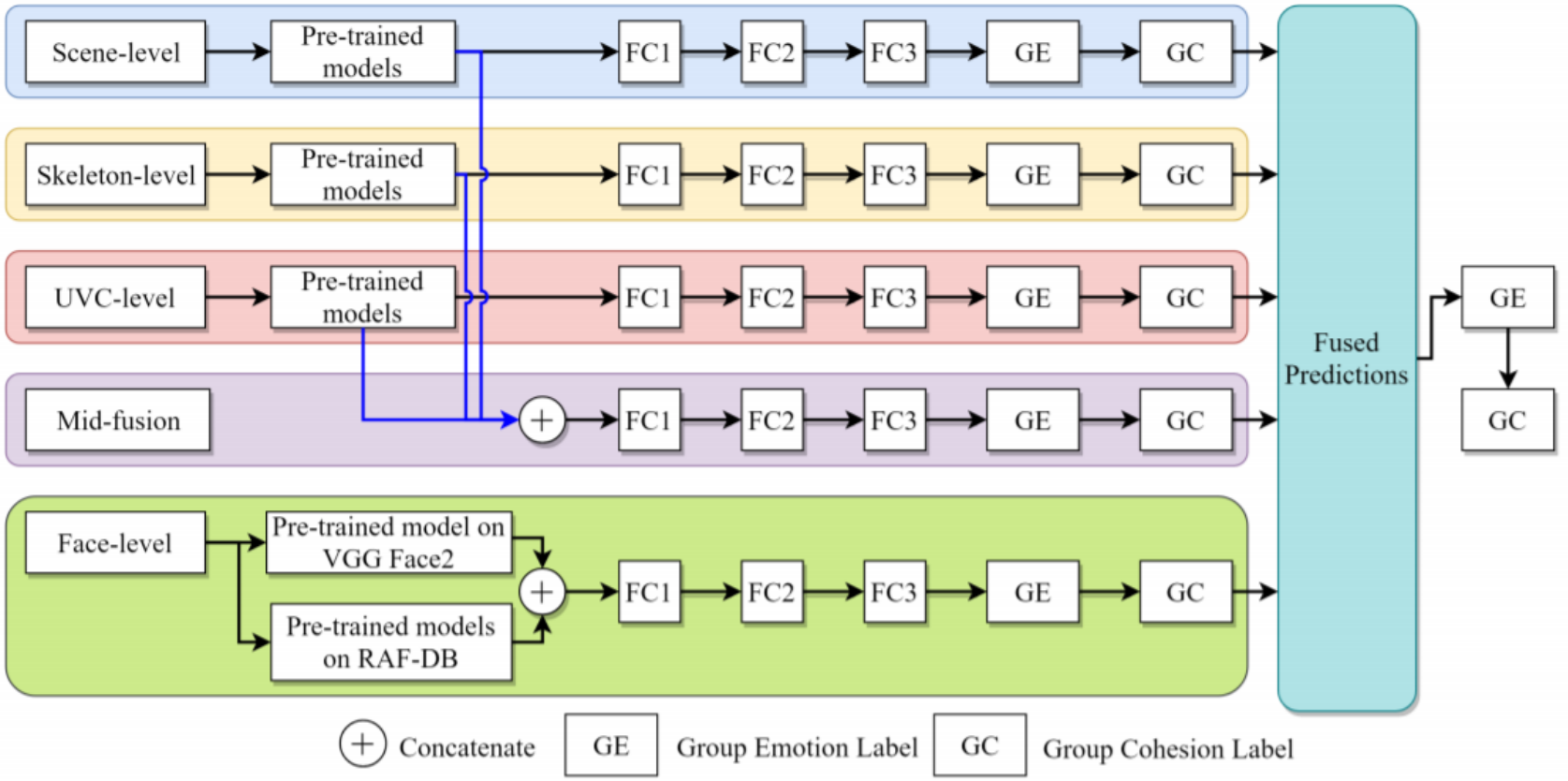
**MSE = 0.56**




Guo, D., Wang, K., Yang, J., Zhang, K., Peng, X., Qiao, Y.: Exploring regularizations with face, body and image cues for group cohesion prediction. In: Proceedings of the International Conference on Multimodal Interaction (ICMI). pp. 557{561. ACM (2019)

# REVIEW

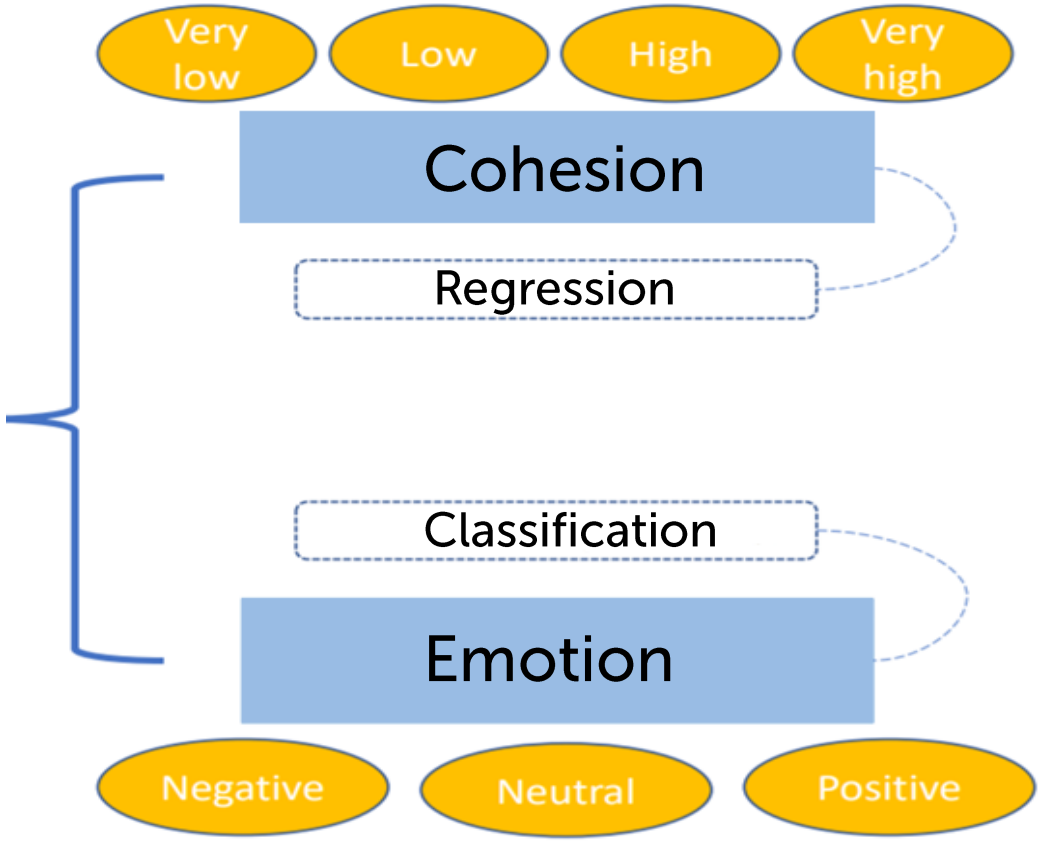
**MSE = 0.52**




Xuan Dang, T., Kim, S.H., Yang, H.J., Lee, G.S., Vo, T.H.: Group-level cohesion prediction using deep learning models with a multi stream hybrid network. In: Proceedings of the International Conference on Multimodal Interaction (ICMI). pp. 572{576. ACM (2019)

# MODEL

Input


Very low
Low
High
Very high
Cohesion
Regression
Classification
Emotion
Negative
Neutral
Positive

# ALGORITHM

**Algorithm 1** Proposed procedure of video-based cohesion prediction.

**Require:** Video frames or images $\{X(t)\}, t = 1, 2, ..., T$
**Ensure:** Cohesion and Emotion label of the given video

1: **for** each frame $t = 1, 2, ..., T$ **do**
2: Obtain $R \geq 0$ facial regions using, e.g., MTCNN face detector
3: **for** each facial area $r = 1, 2, ..., R$ **do**
4: Extract embeddings and simultaneously predict age and gender using the multi-output MobileNet [10]
5: Concatenate embeddings and predicted age and estimate of male gender posterior probability into a single descriptor $\mathbf{x}_r(t)$
6: **end for**
7: Compute the frame feature vector $\mathbf{x}(t)$ as an average of embeddings $\{\mathbf{x}_r(t)\}, r = 1, 2, ..., R$ for all facial regions and normalize it
8: Feed the features into the multi-output neural network
9: Assign the vector of scores $s_{c;cohesion}(t)$ and $s_{c;emotion}(t)$ from the output of regression and classification layers for cohesion and emotion prediction, respectively
10: **end for**
11: Compute the cohesion scores $s_{c;cohesion}$ as an average of scores $\{s_{c;cohesion}(t)\}, t = 1, 2, ..., T$ for all frames
12: Compute the group-level emotion scores $s_{c;emotion}$ as an average of scores $\{s_{c;emotion}(t)\}, t = 1, 2, ..., T$ for all frames
13: Return the cohesion and group emotion categories with the maximal scores $\text{argmax} s_{c;cohesion}$ and $\text{argmax} s_{c;emotion}$.

# PIPELINE

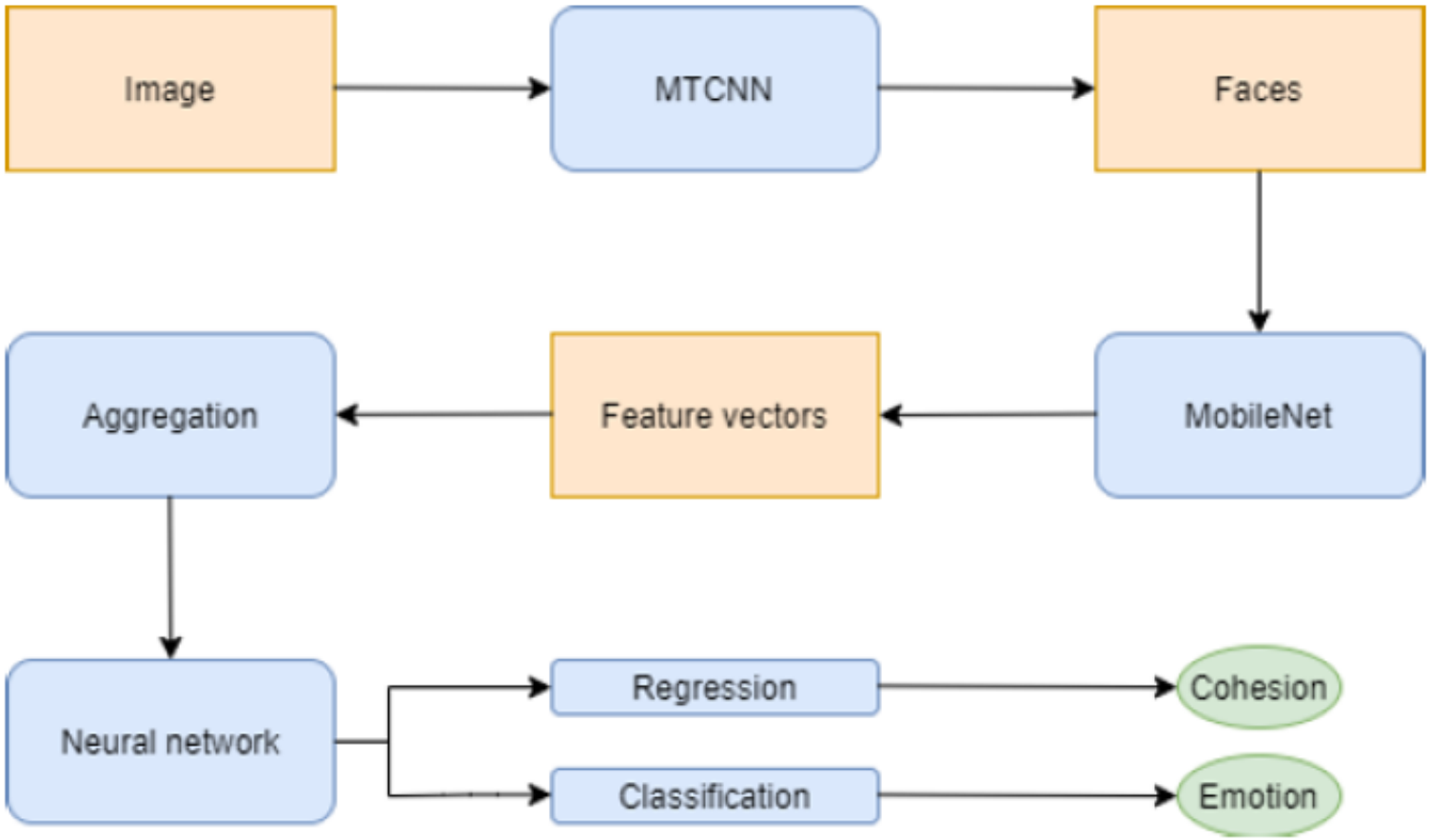
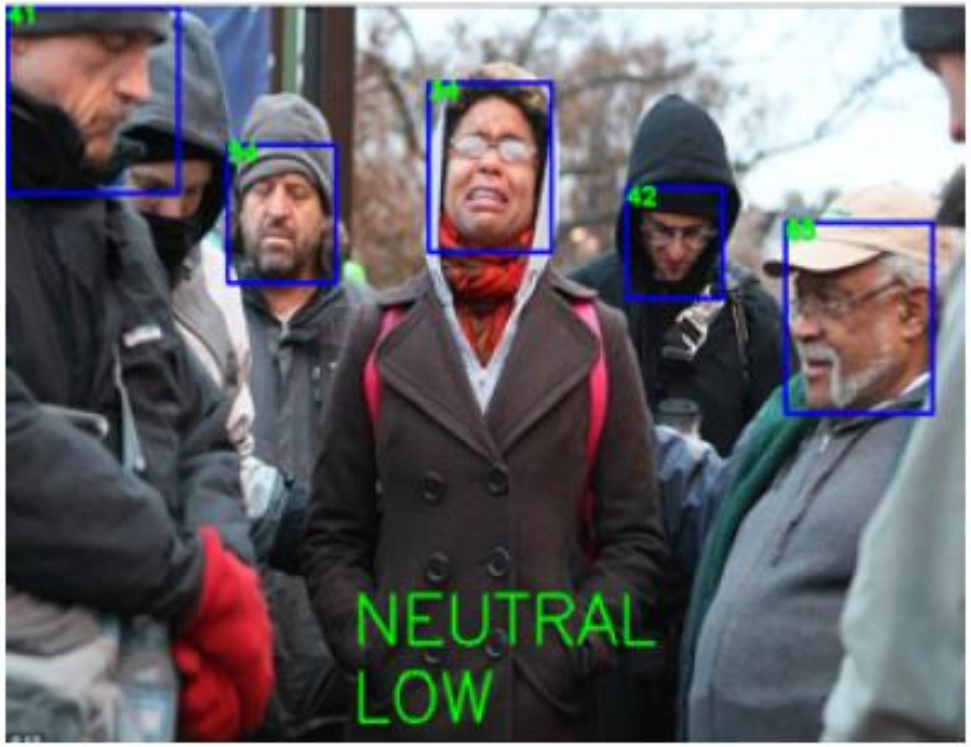



Fig. 3: Proposed pipeline

# APPLICATION


(a)




(b)

# EXPERIMENTS

Table 1: Results for group cohesion prediction, VGGFace2 facial features

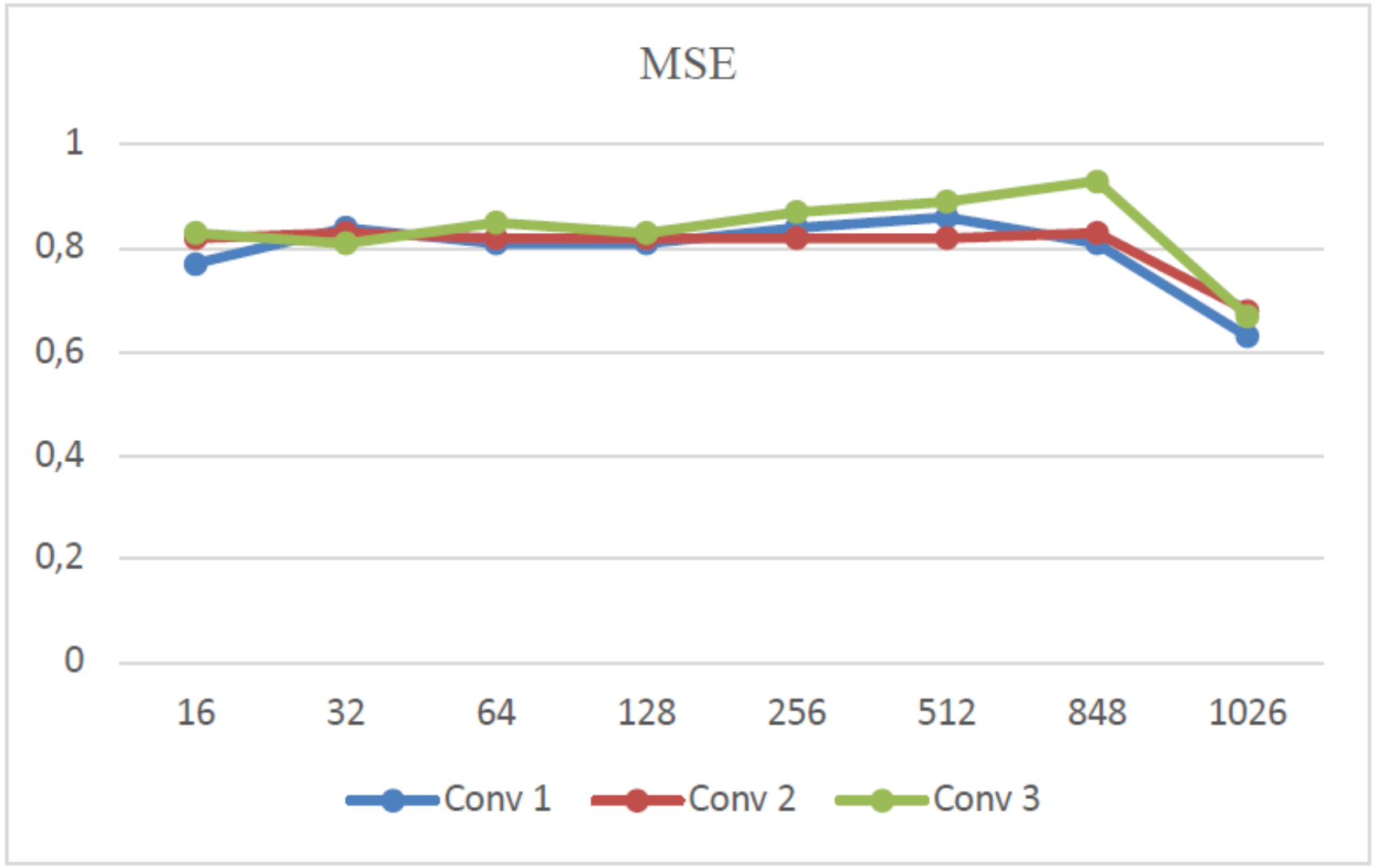
| Supervised learning methods | MSE |
|---|---|
| Ordinal Ridge [16] | 1.01 |
| Ordinal Ridge Regularized [16] | 0.92 |
| Logistic All Threshold [16] | 1.07 |
| Logistic All Threshold Regularized [16] | 0.87 |
| Logistic Immediate Threshold [16] | 1.20 |
| Logistic Immediate Threshold Regularized [16] | 0.97 |
| Logistic SE (Squared Error) [16] | 1.06 |
| Logistic Squared Error Regularized [16] | 0.85 |
| Multi-class Logistic Regression | 1.03 |
| Least Absolute Deviation | 1.05 |
| Catboost [17] | 0.96 |
| Catboost-based [17] Ordinal Classifier | 0.87 |

# EXPERIMENTS

Table 2: Results for facial descriptors for group cohesion prediction, LogisticSE

| Facial descriptor | MSE | Feature extraction time, ms. CPU | GPU |
|---|---|---|---|
| VGGFace (VGG-16) [15] | 1.12 | 109.74 | 8.61 |
| VGGFace2 (ResNet-50) [14] | 0.85 | 57.14 | 11.54 |
| Multi-output MobileNet [10] | 0.80 | 19.94 | 4.76 |

# EXPERIMENTS



# CONCLUSION

- Achieved MSE of cohesiveness prediction on a validation set is 0.21 lower when compared to MSE (0.84) of the baseline from the EmotiW2019.
- The proposed approach is implemented in a publicly-available demo application. In this demo, we predict age and gender of each person and predict the cohesiveness and emotion of the whole group.
- It is rather fast (+ 10 FPS for at most 16 persons in a group using Nvidia GTX1080 Ti GPU) due to the usage of MobileNet. Our preliminary results demonstrated that our model can be used even at Android mobile device with 5 FPS for a small group of 3 persons.
- Our approach is not obviously the best one, as our MSE on validation set is 0.11 and 0.07 greater than MSEs of the rst [6] and second [7] places in the EmotiW 2019 challenge. However, their running time is much worth: slower than 0.35 and 0.2 FPS for ensembles from [6] and [7], respectively.

# FUTURE WORK

- Unfortunately, our group-level emotion recognition accuracy is rather low (0.69), so that it is necessary to further improve our model. Moreover, in future, it is important to extract the faces from a group photo, which significantly influence the overall cohesiveness score.
- It is necessary to examine the state-of-the-art face detectors, e.g., RetinaFace [13], instead of MTCNN in order to locate more faces accurately. However, it is still possible that better facial detectors won't lead to the better quality of cohesiveness prediction, because very small faces do not have robust facial features.

# THANK YOU FOR ATTENTION