Learning to Fuse Latent Representations for Multimodal Data

Oyebade Oyedotun, Aouada Djamila, Björn Ottersten
SnT, University of Luxembourg

SIGCOM/Computer Vision Group

Abstract

Multimodal learning leverages data from different modalities to improve model performance. However, the informativeness of the different data modalities can easily vary across a collected dataset. As such, naively (directly) fusing the latent representations obtained for different modalities may burden the model in finding concise representations that are indeed useful for learning. In this paper, we propose to instead learn the fusion of latent representations for multimodal data by using a modality gating mechanism that allows the dynamic weighting of extracted latent representations based on their informativness. Experiments using the BU-3DFE facial expression recognition and the Washington RGB-D object classification datasets show that learning the fusion of the latent representations for different data modalities leads to improved model generalization.

Background and problem statement

- Let F^T be the factors of variations in a hypothetical dataset, T
- Suppose that we have multimodal data A and data B such that

→ $A \subset T$ and $B \subset T$ for F^A and F^B, respectively as in $F^D = \{F^A, F^B\}$

- Let useful F be F_u^A and F_u^B : $\exists\, F_u^A \subset F^A$ and $F_u^B \subset F^B$
- Learning problem addressed: $| F_u^A | \gg | F_u^B |$ or $| F_u^B | \gg | F_u^A |$

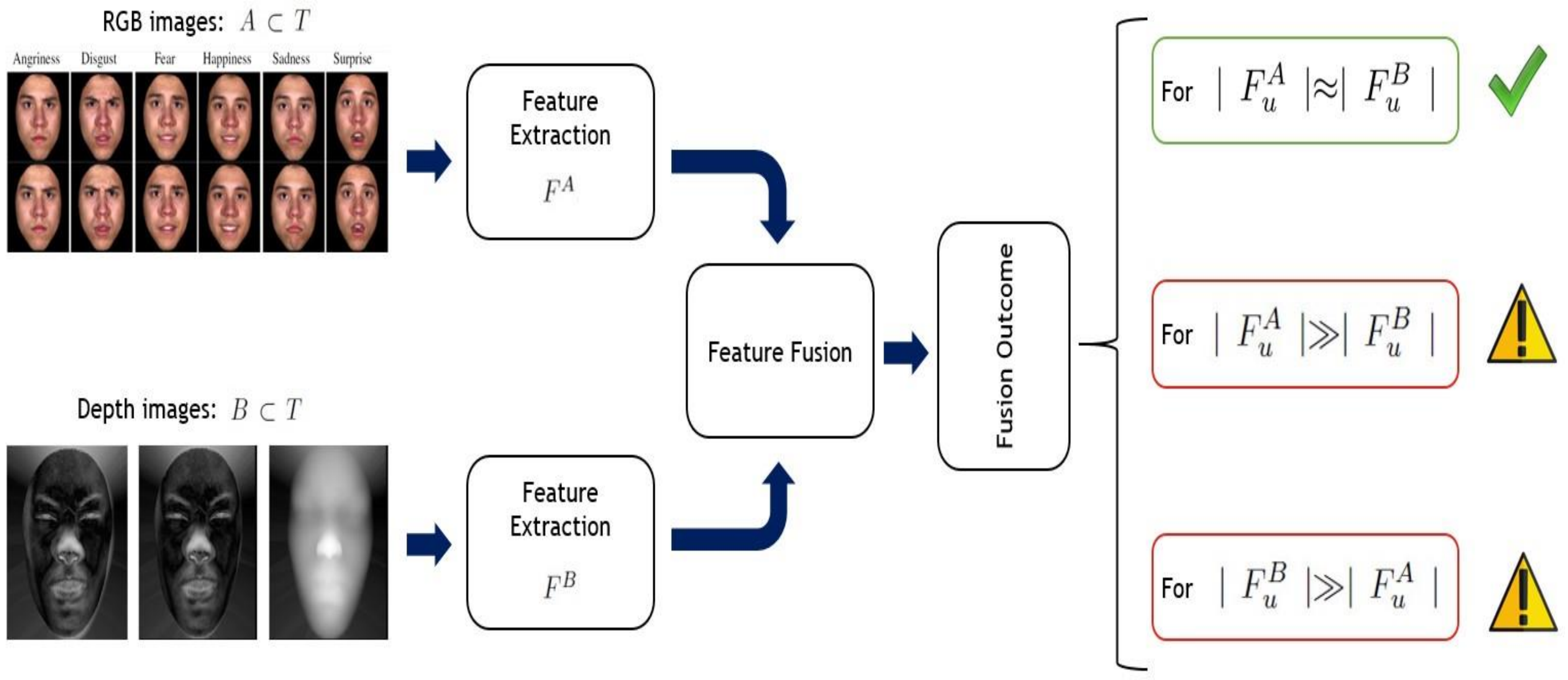
RGB images: $A \subset T$
Angriness
Disgust
Fear
Happiness
Sadness
Surprise
Feature Extraction
F^A
Feature Fusion
Fusion Outcome
Depth images: $B \subset T$
Feature Extraction
F^B
For $| F_u^A | \approx | F_u^B |$
For $| F_u^A | \gg | F_u^B |$
For $| F_u^B | \gg | F_u^A |$

Fig.1. Direct fusion using RGB & Depth images from BU-3FE dataset

Proposed approach and contribution

- Gate transformation, G, on F given $G = \varphi(WF^B + \theta)$ [1]; where transformation on F using weight W, bias θ and Sigmoid function φ
- Fusion outcome is given as $F^O = \{F^A(1-(G(F^B)) \otimes F^B(G(F^B)\}$, where $\otimes$ is the concatenation operation

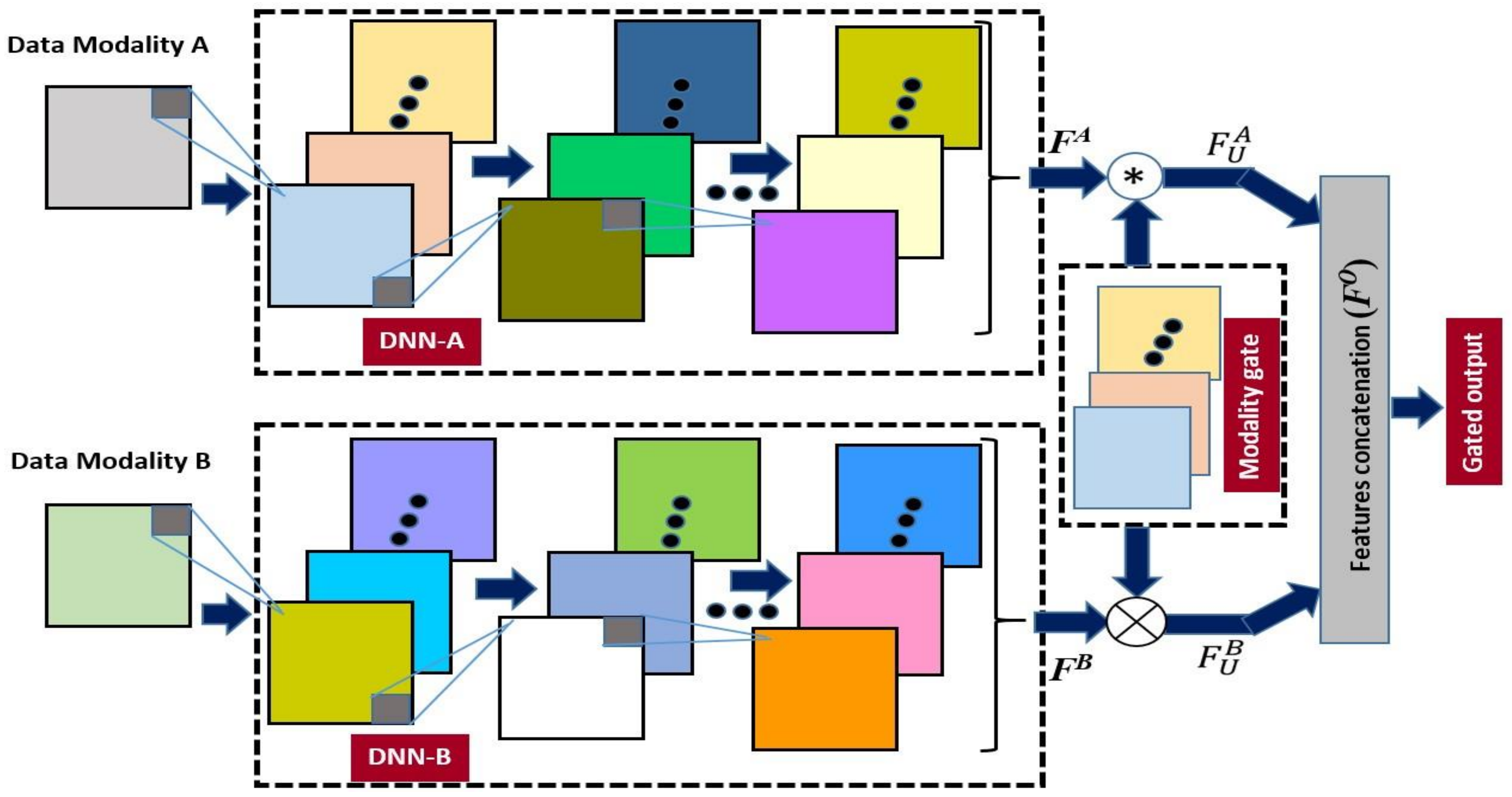

Fig.2. Proposed data dependent fusion of latent representations, F^A and F^B extracted using DNN-A and DNN-B, respectively

Experiments

Table.1. BU-3DFE RGB-D dataset results with 10-fold CV

Approach	Test acc. (%)
DepthMap-ResNet50 [14]	61.11
DepthMap-VGG19 [14]	28.06
DepthMap-scratch [14]	84.72
RGB-ResNet50 [14]	82.92
RGB-VGG19 [14]	81.25
NF: RGB-ResNet50+DepthMap-scratch	87.08
NF: RGB-VGG19+DepthMap-scratch	89.31
Ours: RGB-ResNet50+DepthMap-scratch	**89.86**
Ours: RGB-VGG19+DepthMap-scratch	**90.69**

Table.2. Results comparison on BU-3DFE dataset

Approach	Test acc. (%)
3D geometric shape model+LDA [15]	83.60
Bayesian Belief net+statistical facial features [16]	82.30
Distance+slopes+SVM [18]	87.10
2D+3D features fusion+SVM [19]	86.32
Geometric scattering representation+SVM [20]	84.80
Geometric+photometric attributes+VGG19 [21]	84.87
NF:RGB-ResNet50+DepthMap-scratch [14]	87.08
NF: RGB-VGG19+DepthMap-scratch [14]	89.31
Ours: RGB-ResNet50+DepthMap-scratch	**89.86**
Ours: RGB-VGG19+DepthMap-scratch	**90.69**

Table.3. Washington RGB-D dataset results with 5-fold CV

Approach	Test acc. (%)
Depth map-ResNet50	1.67
Depth map-VGG19	2.38
DepthMap-scratch	73.87
RGB-ResNet50	29.76
RGB-VGG19	96.73
NF: RGB-ResNet50+DepthMap-scratch	61.23
NF: RGB-VGG19+DepthMap-scratch	83.16
Ours: RGB-ResNet50+DepthMap-scratch	**98.93**
Ours: RGB-VGG19+DepthMap-scratch	**99.51**

Results on BU-3DFE facial expression [2] and Washington RGB-D object classification datasets are given in Tables 1, 2 & 3.

- For RGB data, pre-trained VGG-16 and ResNet50 are used.
- For depth maps, training a small from scratch performs better.
- Learning the fusion of latent representations improves results.

Conclusion

The latent representations extracted from the different modalities are typically fused via naïve fusion (i.e. direct concatenation); we show that this can even hurt model performance. Instead, we propose to allow the model to learn a data driven fusion stage using a gate mechanism that filters latent representations from multimodal data. Experimental results on two different datasets validate the proposed fusion approach.

References

[1] Oyedotun, O. K., Aouada, D., & Ottersten, B. (2019, May). Learning to Fuse Latent Representations for Multimodal Data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3122-3126). IEEE.

[2] Oyedotun, O. K., Abd El Rahman Shabayek, A., Aouada, D., & Ottersten, B. (June, 2018). Highway Network Block with Gates Constraints for Training Very Deep Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1658-1667).

[3] Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006, April). A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)* (pp. 211-216). IEEE.

This work was funded by:
The National Research Fund (**FNR**), Luxembourg, under the project reference R-AGR- 0424-05-D/Bjorn Ottersten, and FNR project IDform under the agreement CPPP17/ IS/11643091/IDform/Aouada.

Fonds National de la Recherche Luxembourg