Towards Visually Prompted Keyword Localisation For Zero-Resource Spoken Languages

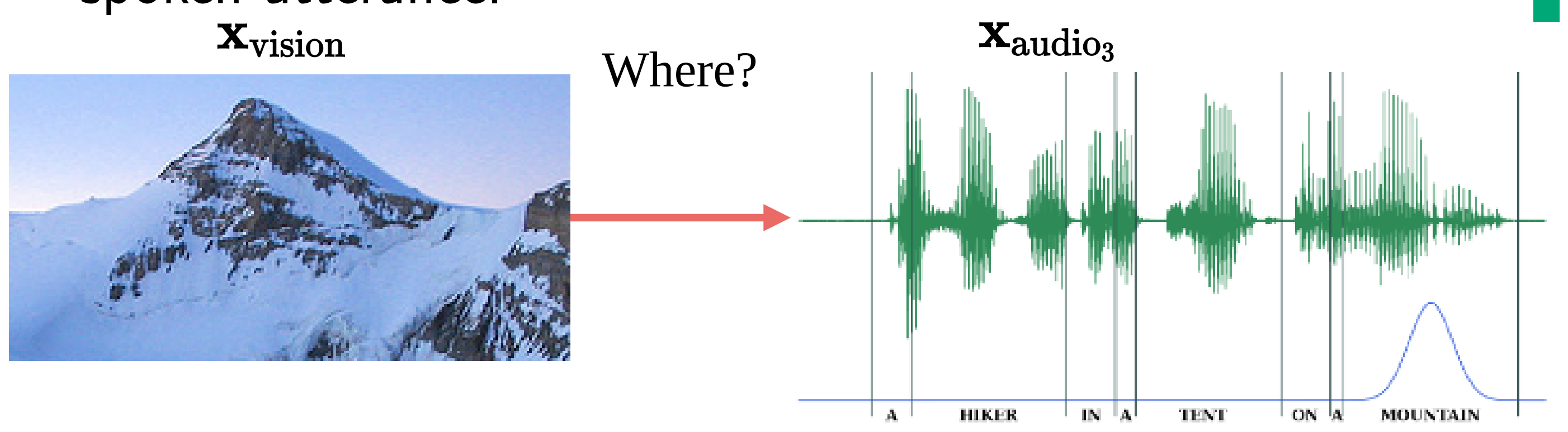
Leanne Nortje & Herman Kamper

nortjeleanne@gmail.com | kamperh@sun.ac.za

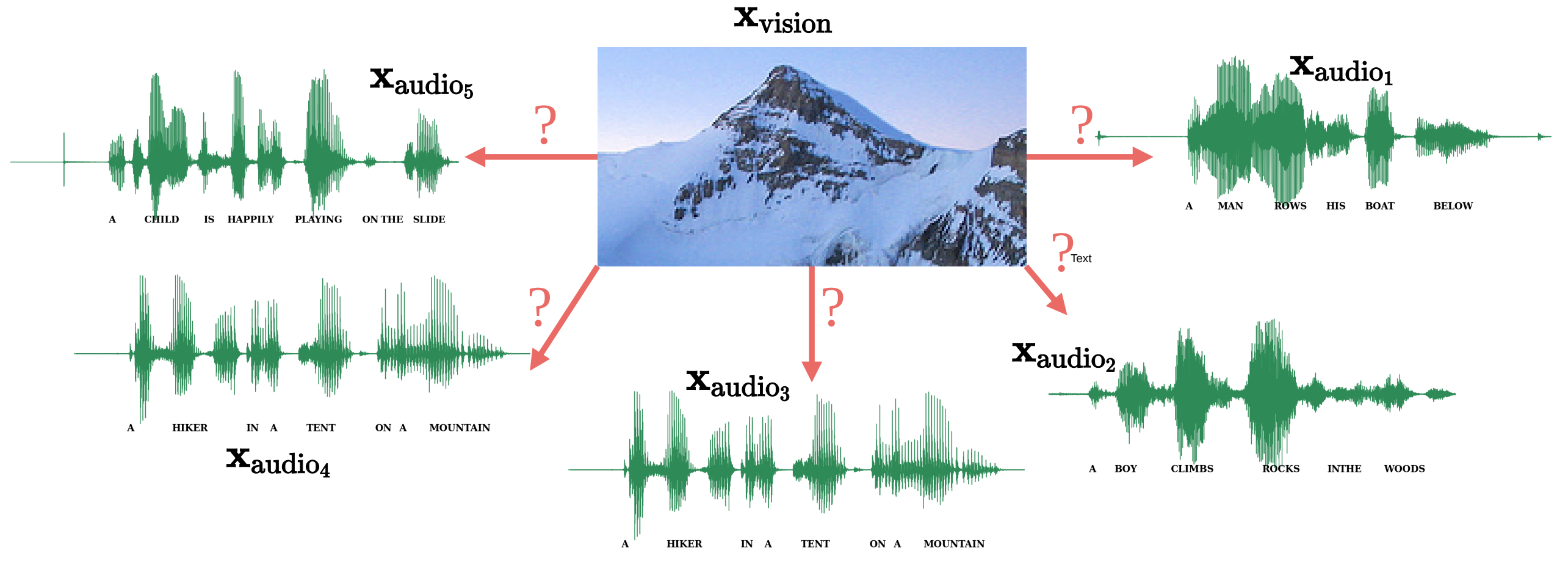
Electrical & Electronic Engineering, Stellenbosch University, South Africa

1. The Task

- What if a language does not have a written form?
- The goal in visually prompted keyword localisation is to locate a given query keyword (given as an image) within a spoken utterance.

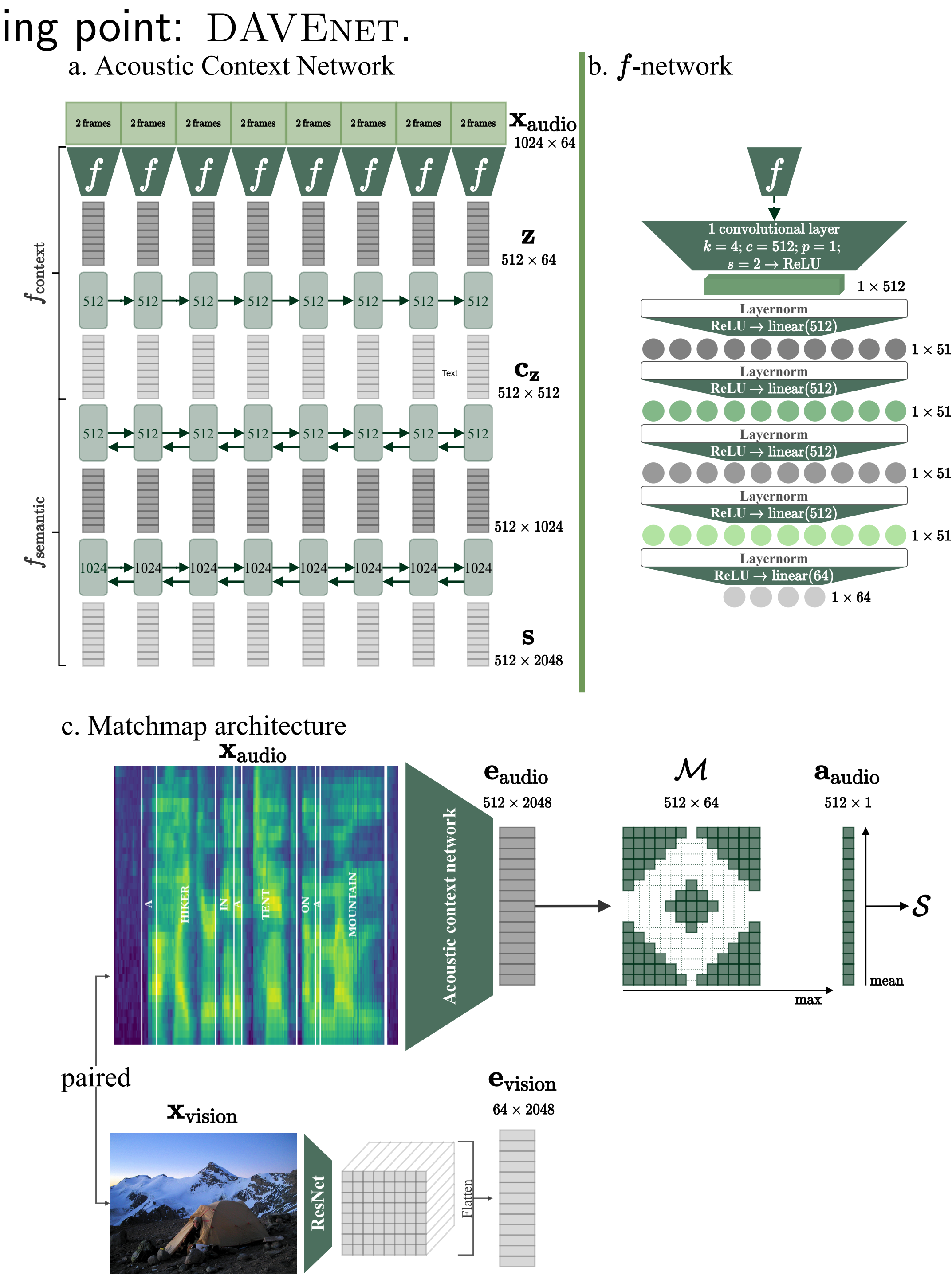


- The goal in visually prompted detection is to detect whether a given query keyword (given as an image) occurs anywhere within a spoken utterance.


$\mathbf{x}_{\text{audio}_5}$
$\mathbf{x}_{\text{vision}}$
$\mathbf{x}_{\text{audio}_1}$
$\mathbf{x}_{\text{audio}_2}$
$\mathbf{x}_{\text{audio}_3}$
$\mathbf{x}_{\text{audio}_4}$
?
A CHILD IS HAPPILY PLAYING ON THE SLIDE
A HIKER IN A TENT ON A MOUNTAIN
A BOY CLIMBS ROCKS IN THE WOODS

2. The Models

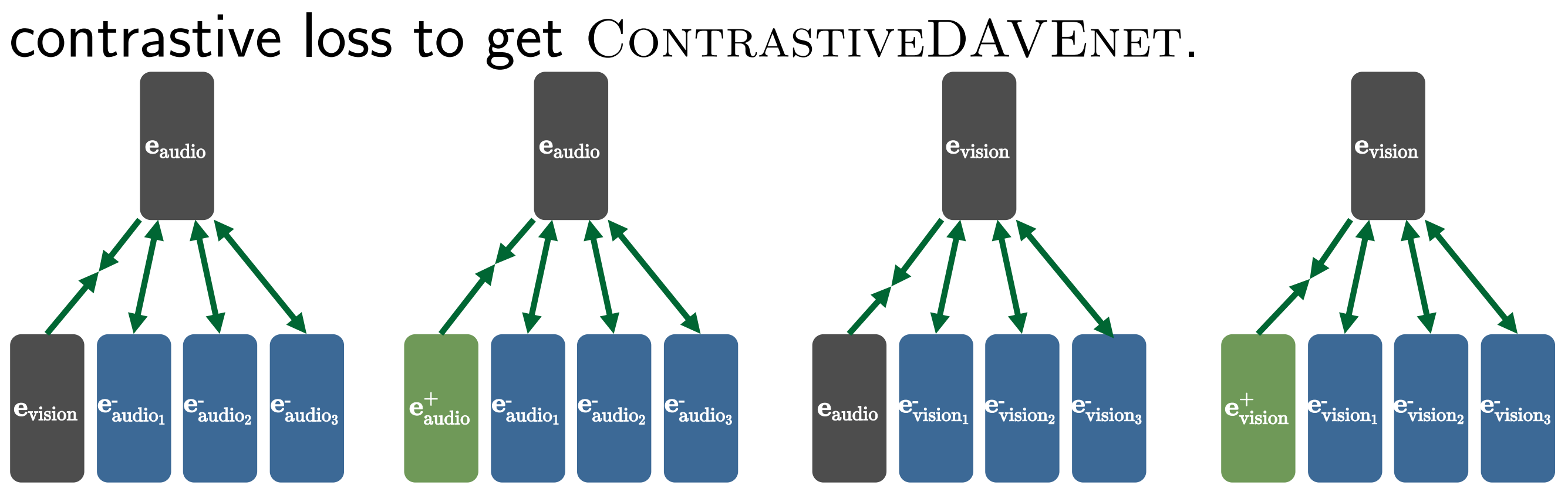
- Starting point: DAVEnet.


a. Acoustic Context Network

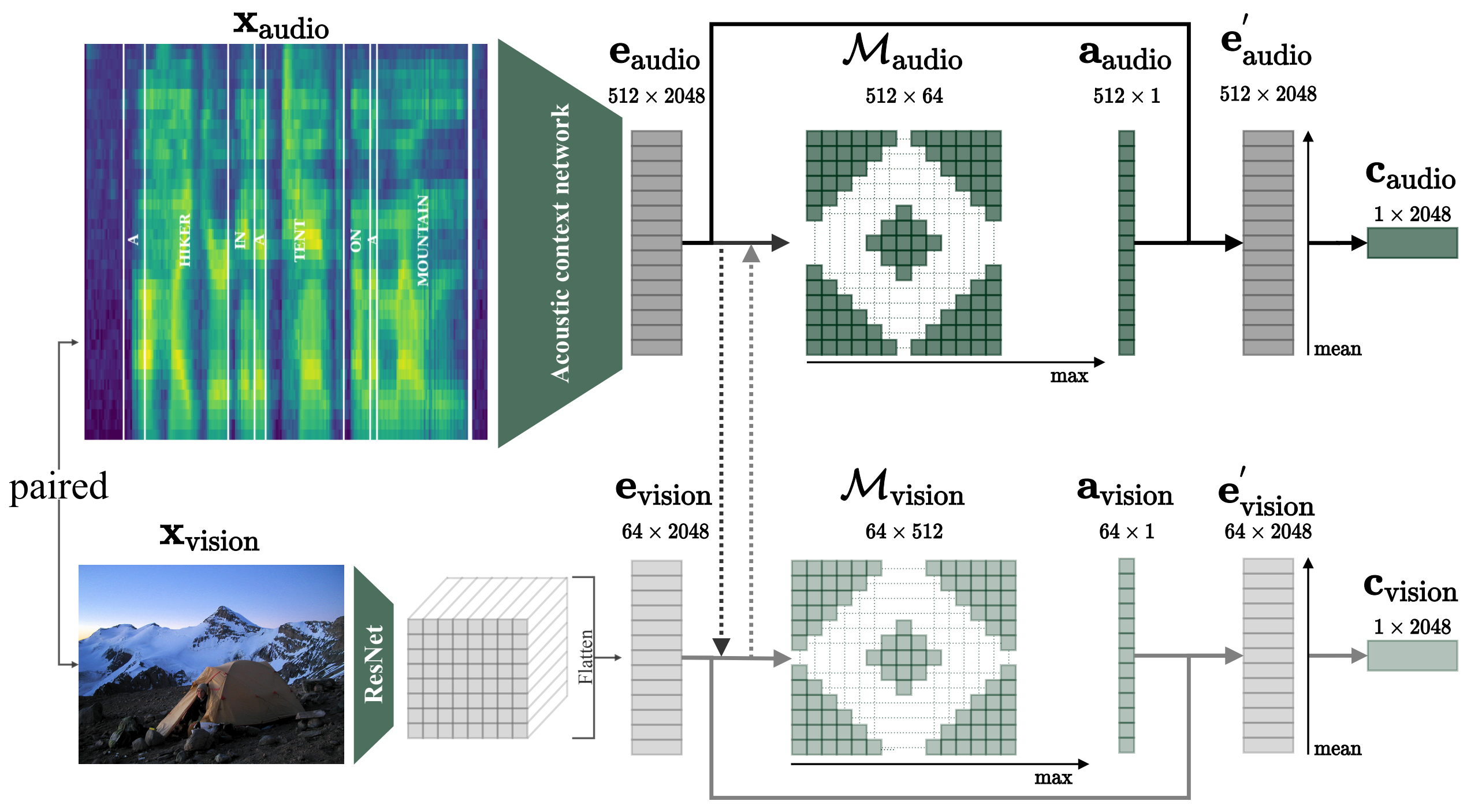
b. f-network

c. Matchmap architecture

- We add positive and negative samples to DAVEnet with a contrastive loss to get ContrastiveDAVEnet.


$\mathbf{e}_{\text{audio}}$
$\mathbf{e}_{\text{vision}}$
$\mathbf{e}^{-}_{\text{audio}_1}$
$\mathbf{e}^{-}_{\text{audio}_2}$
$\mathbf{e}^{-}_{\text{audio}_3}$
$\mathbf{e}^{+}_{\text{audio}}$
$\mathbf{e}^{-}_{\text{vision}_1}$
$\mathbf{e}^{-}_{\text{vision}_2}$
$\mathbf{e}^{-}_{\text{vision}_3}$
$\mathbf{e}^{+}_{\text{vision}}$

- We add a multimodal localising attention mechanism to ContrastiveDAVEnet to get LocalisationAttentionNet.


$\mathbf{x}_{\text{audio}}$
Acoustic context network
$\mathbf{e}_{\text{audio}}$
512 × 2048
$\mathcal{M}_{\text{audio}}$
512 × 64
$\mathbf{a}_{\text{audio}}$
512 × 1
$\mathbf{e}'_{\text{audio}}$
512 × 2048
$\mathbf{c}_{\text{audio}}$
1 × 2048
max
mean
paired
$\mathbf{x}_{\text{vision}}$
ResNet
Flatten
$\mathbf{e}_{\text{vision}}$
64 × 2048
$\mathcal{M}_{\text{vision}}$
64 × 512
$\mathbf{a}_{\text{vision}}$
64 × 1
$\mathbf{e}'_{\text{vision}}$
64 × 2048
$\mathbf{c}_{\text{vision}}$
1 × 2048

3. Baselines

- We randomly sample detection and attention values.
- A visual BoW model which is prompted with written queries instead of image keys.

4. Results

- Keyword detection results (%).

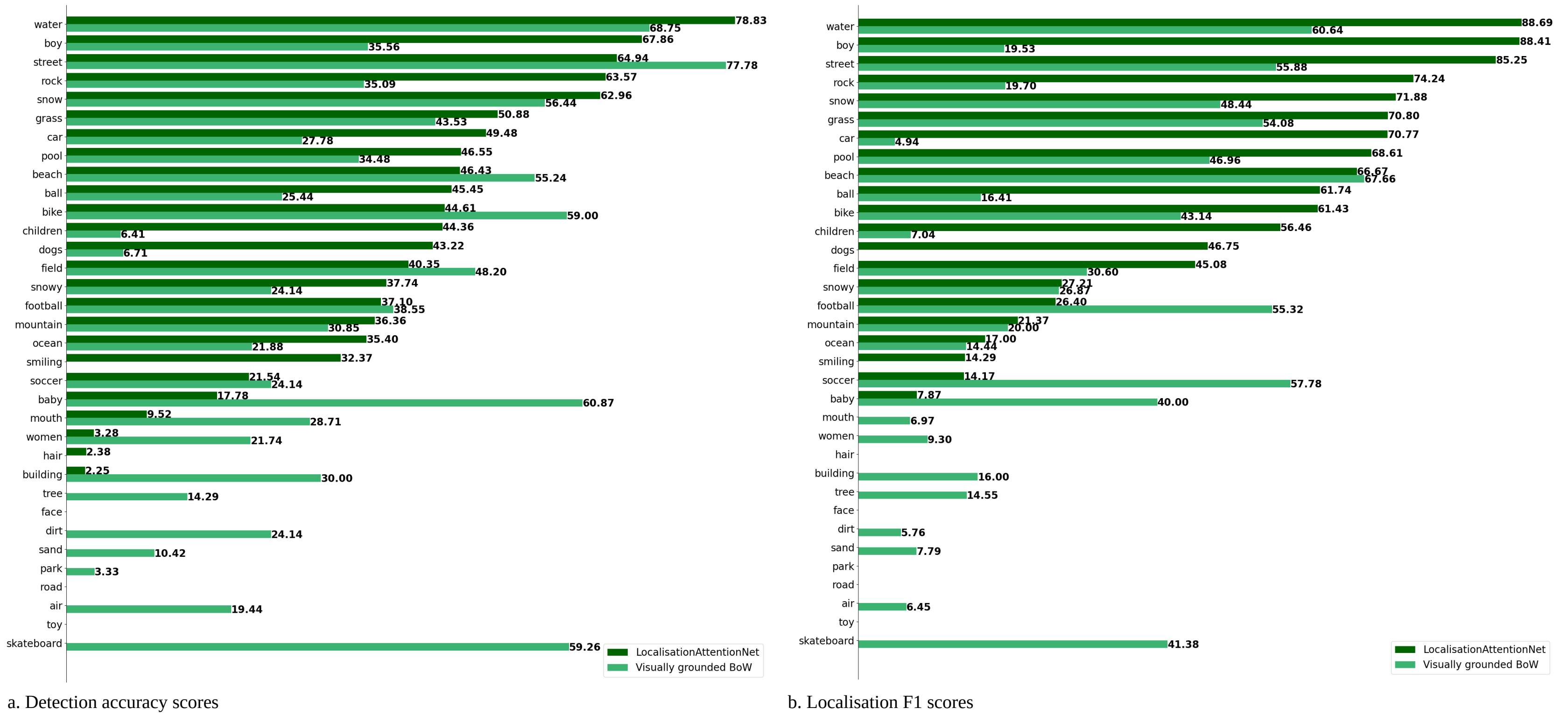
Model	Precision	Recall	F1 score
Text query			
Visually grounded BoW [1]	42.29	36.32	39.08
Image query			
Random baseline	2.30	13.96	3.94
DAVEnet	8.86	46.51	14.88
ContrastiveDAVEnet	37.97	44.84	41.12
LocalisationAttentionNet	**48.41**	**55.85**	**51.86**

- Keyword localisation results (%).

Model	Precision	Recall	F1 score
Text query			
Visually grounded BoW [1]	33.39	31.02	32.17
Image query			
Random baseline	0.13	0.87	0.22
DAVEnet	5.17	33.36	8.95
ContrastiveDAVEnet	30.43	39.45	34.36
LocalisationAttentionNet	**44.43**	**53.77**	**48.66**

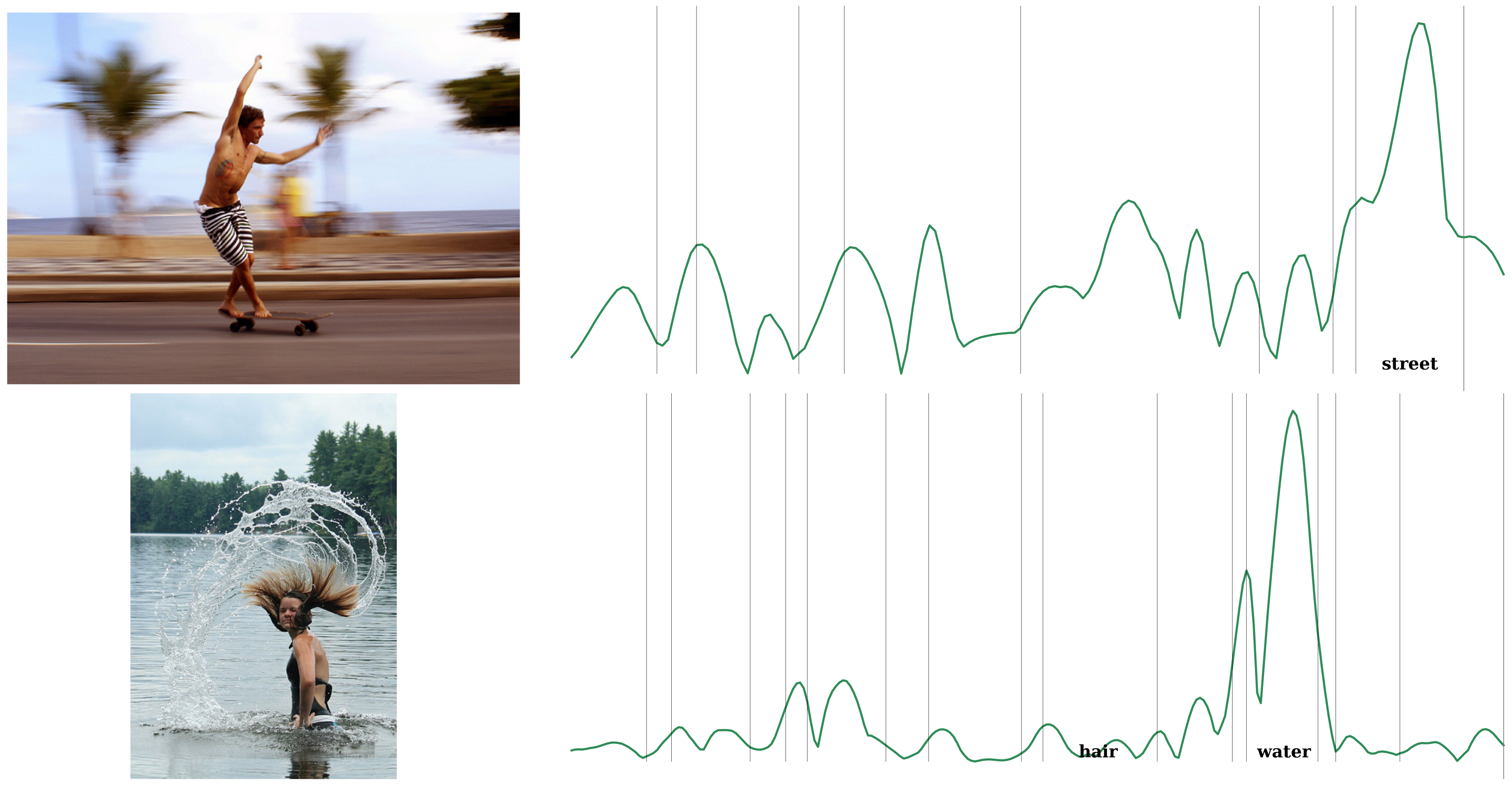
5. Ablation Experiments

- The per keyword (a) detection accuracy scores and (b) localisation F1 score of LocalisationAttentionNet and the visually grounded BoW model.


a. Detection accuracy scores

b. Localisation F1 scores

- The audio attention weights from LocalisationAttentionNet for two utterance-query pairs.



6. Conclusion

- VPKL is more accurate than using word embeddings.
- We need to replace the ideal visual tagger with an actual tagger.

7. References

[1] K. Olaleye, B. van Niekerk, and H. Kamper, "Towards localisation of keywords in speech using weak supervision," in *Proc. NeurIPS-SAS*, 2020.