CoNAN: Conditional Neural Aggregation Network For Unconstrained Face Feature Fusion

Bhavin Jawade*, Deen Dayal Mohan*, Dennis Fedorishin, Srirangaraj Setlur, Venu Govindaraju

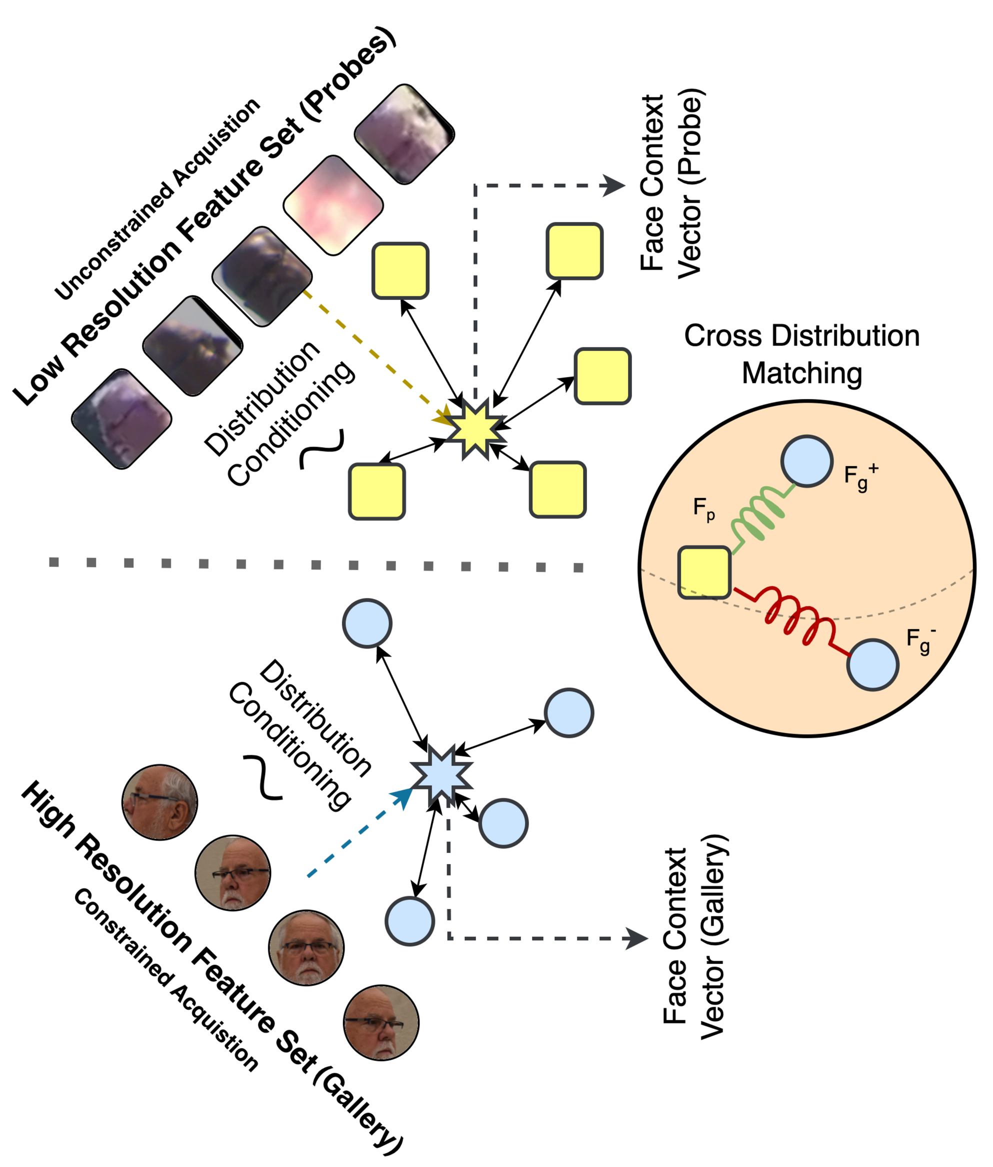
{bhavinja, dmohan, dcfedori, setlur, govind}@buffalo.edu

IEEE International Joint Conference on Biometrics

CUBSLab

Motivation

1. Face recognition in challenging conditions **(e.g., large distances, low resolutions, varying factors)** is difficult.
2. Traditional methods use metadata or high-dimensional features for aggregation, which **may not be practical for low-res, long-range faces.**
3. The proposed **CoNAN approach conditions a context vector on feature set distribution** to weigh features based on **informativeness**.
4. CoNAN **outperforms** existing methods on datasets like BTS and DroneSURF for long-range face recognition.



Problem Statement

An ideal face feature aggregation technique must have the following properties:

1. It should **adapt with varying number of features** in the image-set

2. The method's performance **should not be conditioned on the availability of high-quality metadata** or high dimensional intermediate feature maps from images

3. It should **discount all the non-informative feature** representations and **prioritize highly discriminative** feature embeddings

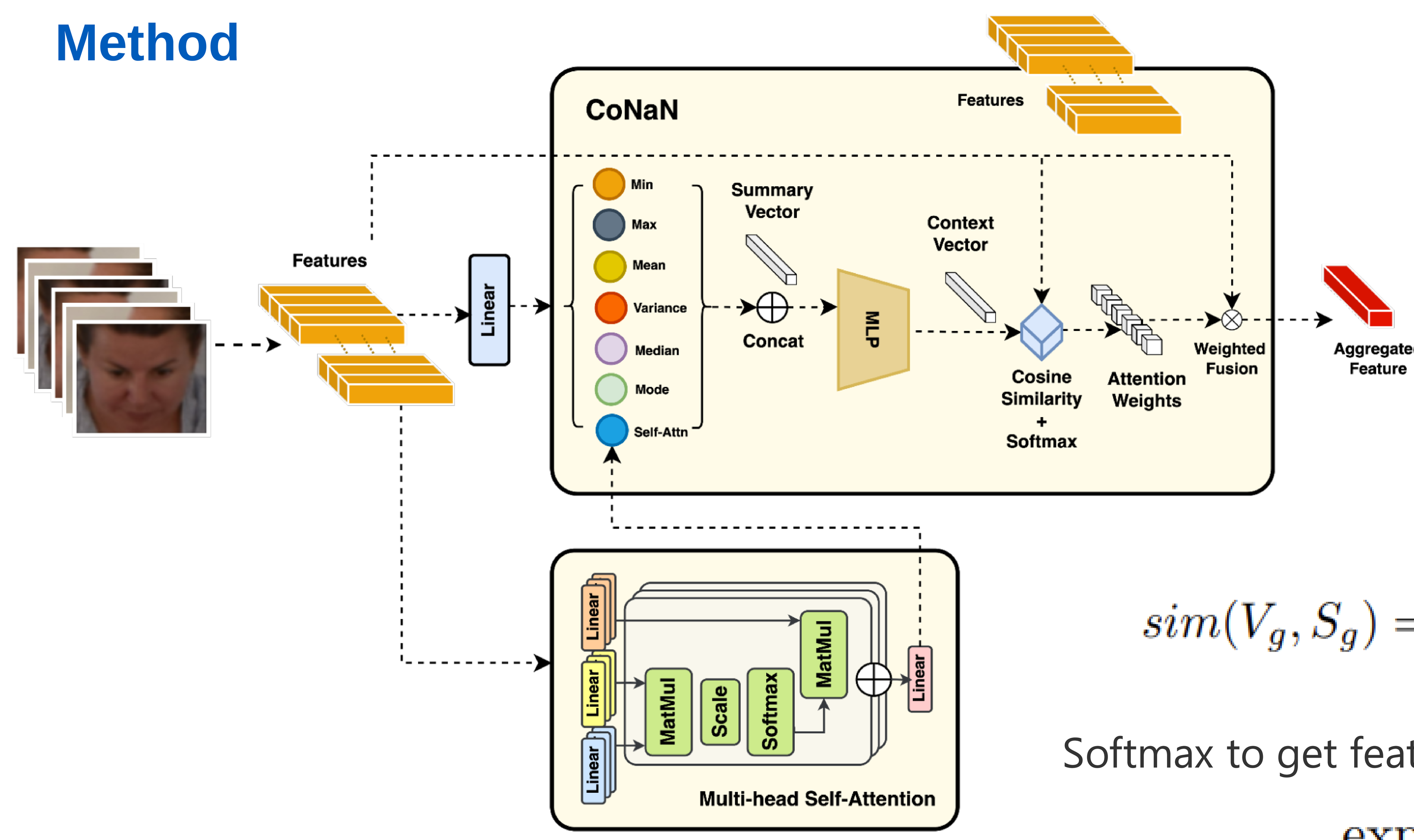
4. It should be able to **adapt to a variety of face feature extractors** with minimal retraining

5. The method should prioritize feature representation in the **gallery that closely matches the distribution of probe** features

6. Should add **minimal computational overhead** to the existing feature representation

* Equal contribution authors

Method



Compute Similarity of features to context vector

$$sim(V_g, S_g) = \frac{1}{N}\sum_{i=1}^{N}\frac{V_g \cdot x_i}{||V_g||\,||x_i||}$$

Softmax to get feature weights

$$Wg_i = \frac{\exp(sim(V_g, S_g)_i/T)}{\sum_{i=1}^{N}\exp(sim(V_g, S_g)_i/T)}$$

Approach

Given feature sets S_g and S_p compute the minimum, maximum , mean, variance, mode, and median along each dimension.

Combine with **Cp or Cg (classification tokens from attention module) and DTE (Learnable Distribution Type Embedding)** - This aids in learning a distribution specific context vector.

$$\vec{g}(S_p) = \{C_p, DTE_p, max(S_p), min(S_p), mean(S_p), var(S_p), mode(S_p), median(S_p)\}$$

$$\vec{g}(S_g) = \{C_g, DTE_g, max(S_g), min(S_g), mean(S_g), var(S_g), mode(S_g), median(S_g)\}$$

Supervised Contrastive Loss

$$\mathcal{L} = -\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(\frac{z_i\cdot z_p}{\tau}\right)}{\sum_{a\in A(i)}\exp\left(\frac{z_i\cdot z_a}{\tau}\right)}$$

Results on **DroneSURF**

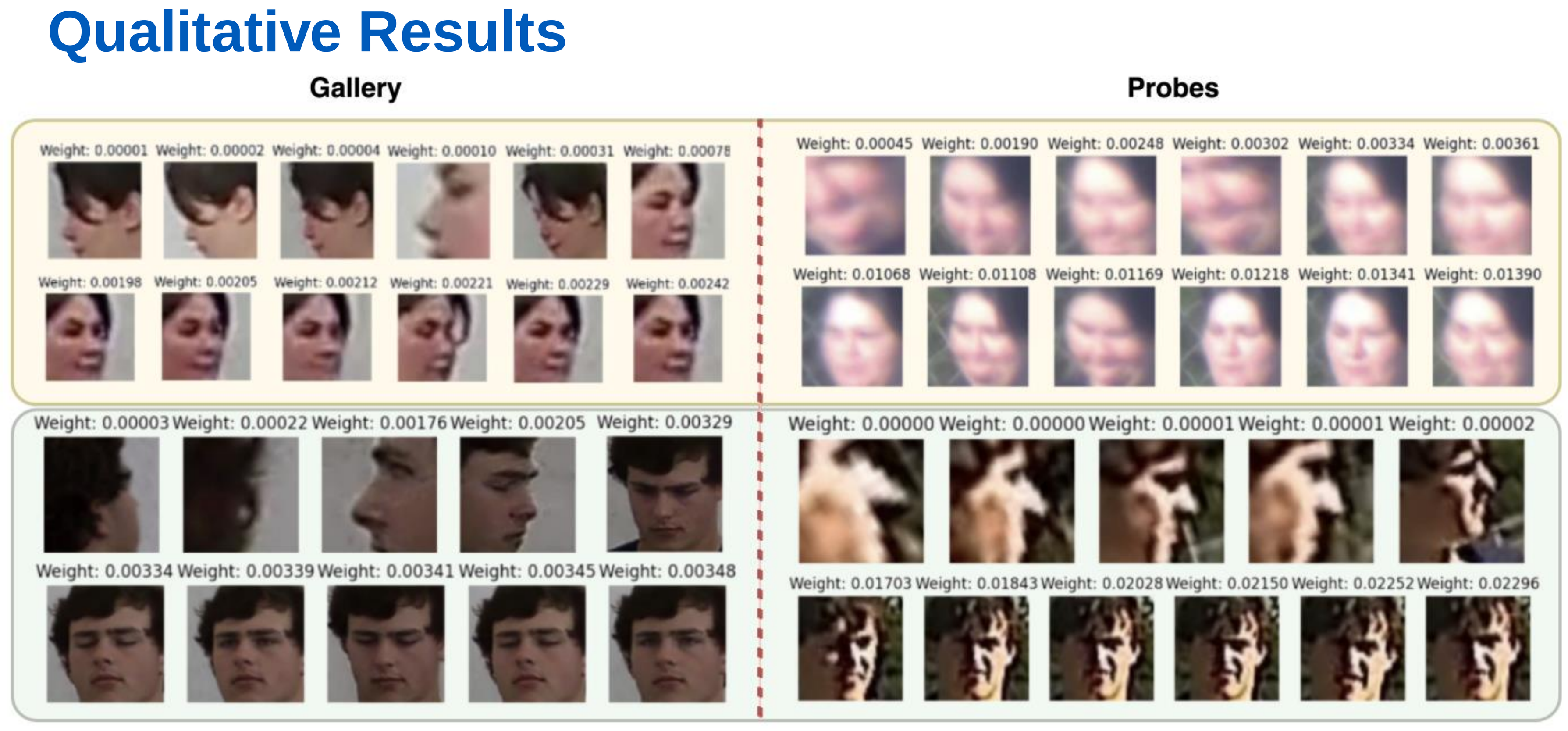
Trained On DroneSURF			
	Feature Extractor	Active	Passive
HOG [5]	-	8.33	7.30
LBP [25]	-	4.16	4.16
VGGFace [26]	-	16.67	5.20
COTS [13]	-	21.88	4.16
GAP [19]	Arcface [6]	16.67	8.33
CoNAN [34]	**Arcface [6]**	**17.71**	**13.54**
GAP [19]	Adaface [15]	46.87	7.29
NAN [34]	Adaface [15]	65.62	6.25
MCN [33]	Adaface [15]	72.92	8.33
CoNAN [34]	**Adaface [15]**	**80.21**	**13.54**

Quantitative Results

Results on **BTS 3.1 (BRIAR)**

		Face Included Treatment				Face Included Control			
	Feature Extractor	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
GAP [19]	Arcface [6]	53.7	37.01	27.28	19.48	91.17	84.82	75.26	66.21
NAN [34]	Arcface [6]	55.41	39.01	26.64	18.3	91.34	84.31	72.9	60.37
MCN [33]	Arcface [6]	55.06	39.41	28.22	19.37	92.41	87.12	77.9	67.17
CoNAN	Arcface [6]	**60.36**	**43.38**	**32.14**	**23.14**	**93.36**	**87.57**	**80.94**	**71.89**
GAP [19]	Adaface [15]	63.79	50.76	40.81	31.7	96.17	91.28	86.9	80.1
NAN [34]	Adaface [15]	65.29	54.44	44.96	34.86	96.06	93.31	90.16	84.82
MCN [33]	Adaface [15]	65.22	54.25	45.01	34.84	96.06	93.19	89.82	85.32
CoNAN	Adaface [15]	**67.56**	**56.32**	**46.14**	**36.52**	**96.06**	**93.7**	**90.27**	**85.72**

Qualitative Results



References:

[SupCon]: Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33 (2020): 18661-18673.
[NAN]: Yang, Jiaolong, et al. "Neural aggregation network for video face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[MCN]: Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192, 2018.