VisionTransformer: Opportunities for Gait Recognition

Qitao Zhao (赵淇涛), 3.2 2022
Advised by Prof. Xianye Ben

Vision Transformer:
Opportunities for Gait Recognition

Content

Introduction to

Gait Recognition

Introduction to Gait Recognition

Charming Properties

- Unique
- Hard to disguise
- Recognized at a distance
- Need no cooperation

Broad Applications

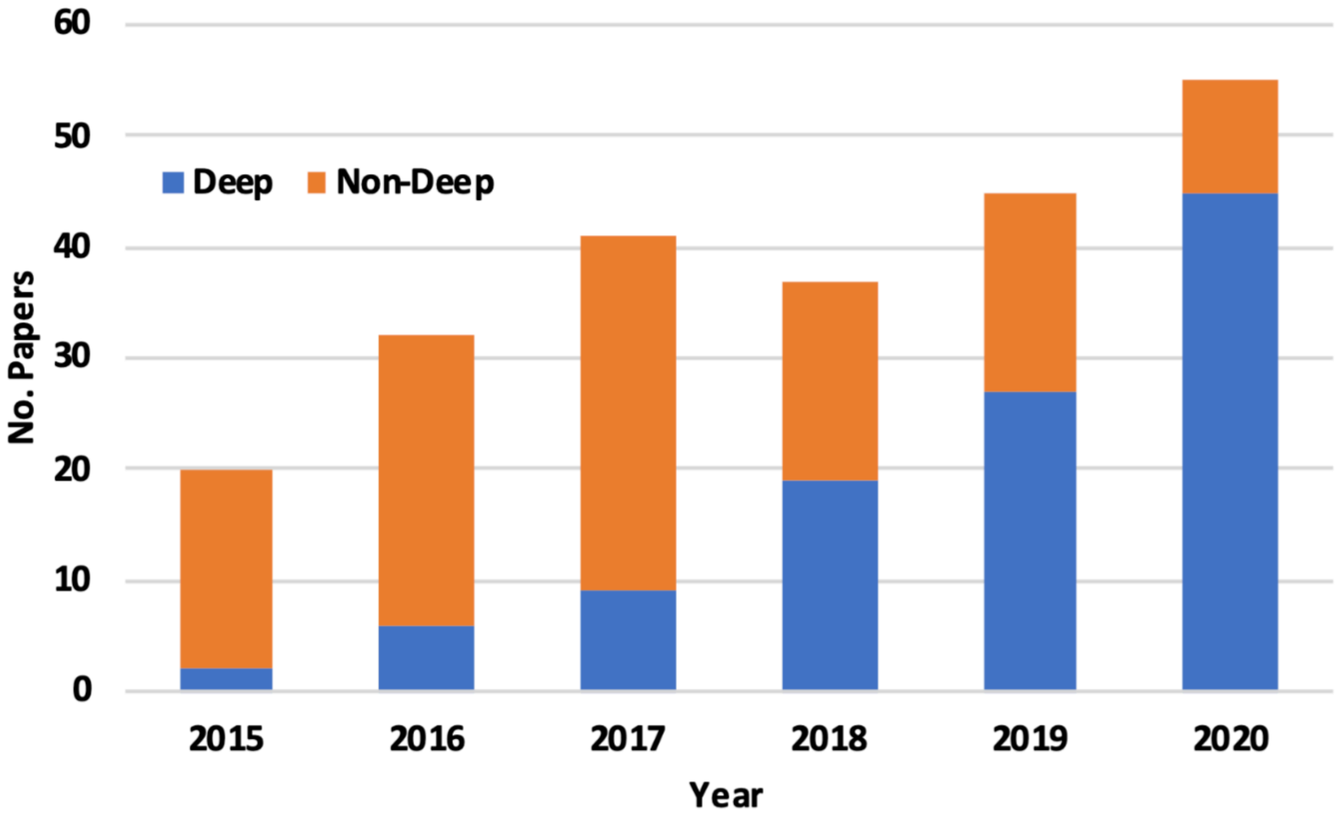
- Individual identification
- Crime prevention
- Forensic identification

A *key* challenge:

Modeling dynamic walking patterns from gait sequence

Existing Methods

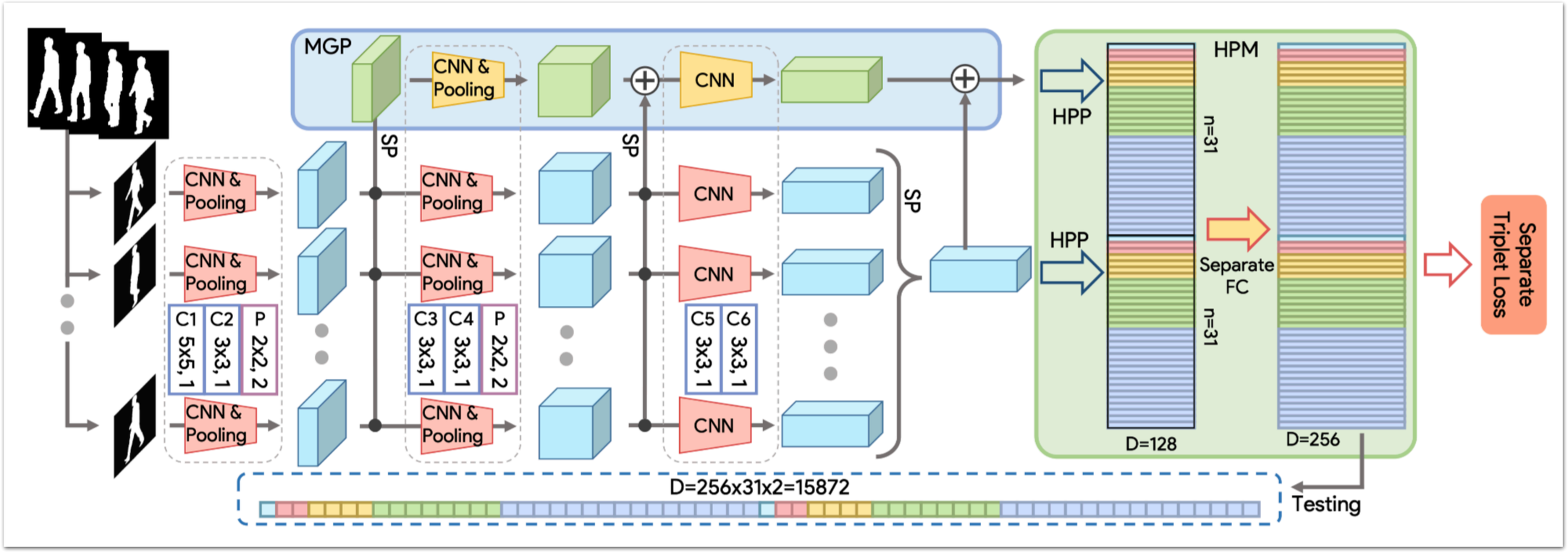
The number of gait recognition papers published after 2015



Sepas-Moghaddam, A., & Etemad, A. (2021). Deep Gait Recognition: A Survey. arXiv preprint arXiv:2102.09546.

Existing Methods

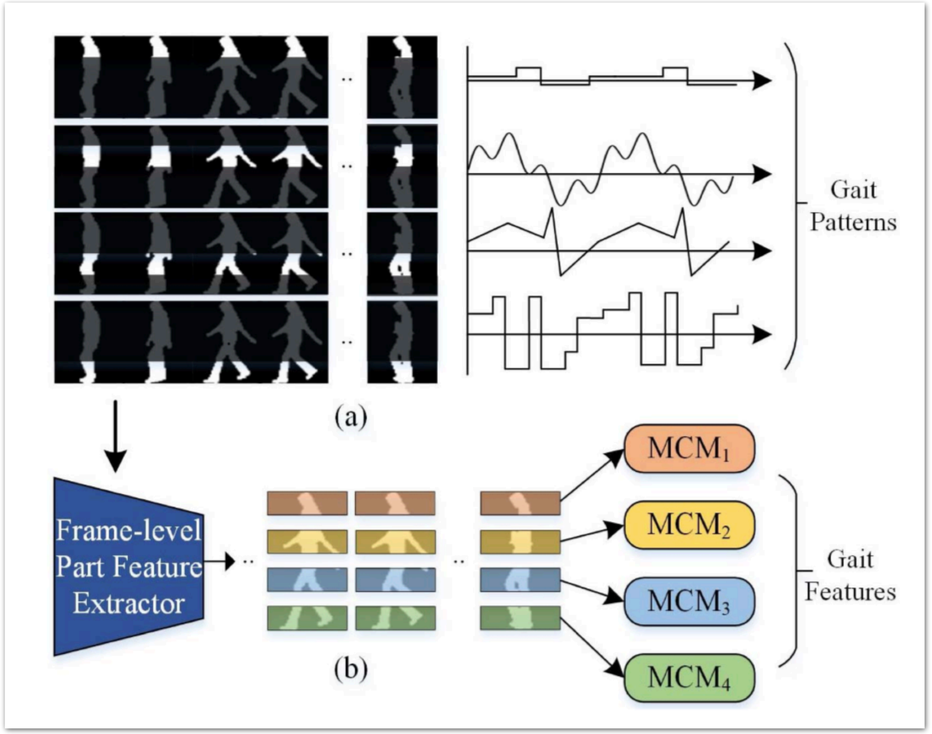
Temporal feature aggregation using *attention* within a non-ordered set

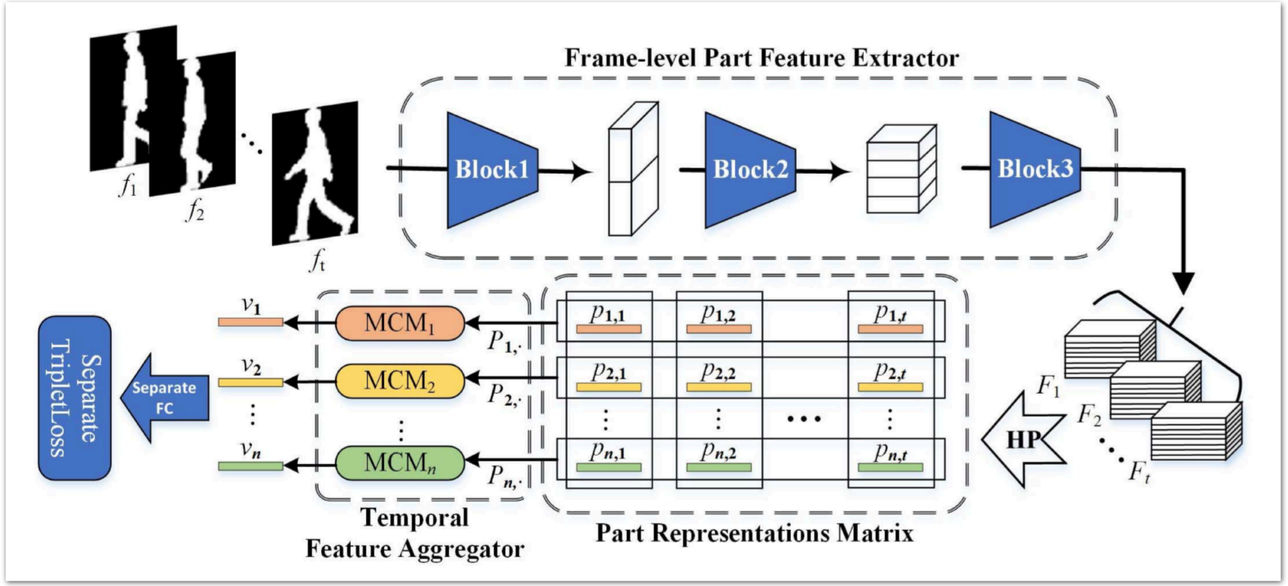


H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in AAAI Conference on Artificial Intelligence, Honolulu, HW, USA, February 2019.

Existing Methods

Temporal feature aggregation using *attention* between neighboring images & Local spatial feature extraction

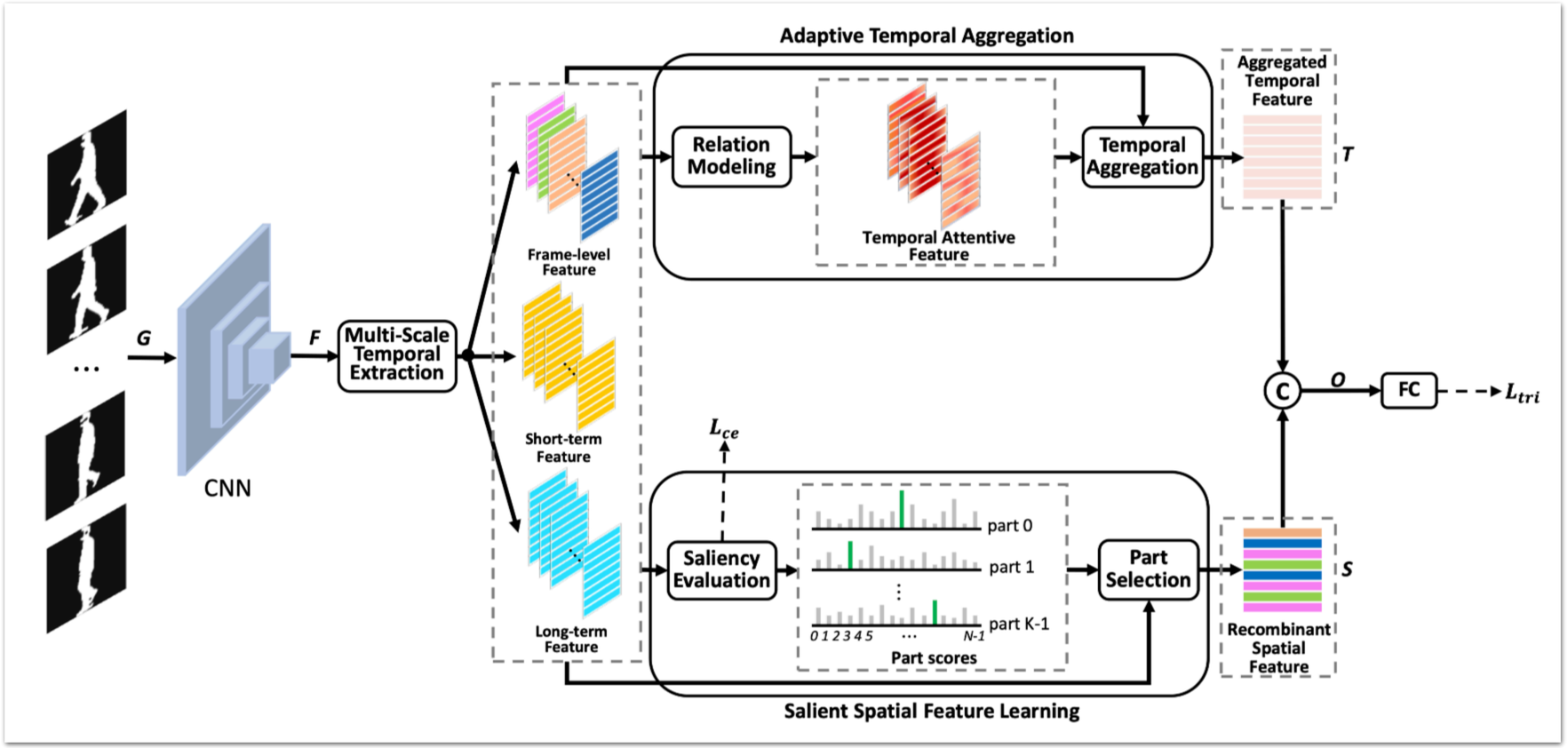




C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition,"in Computer Vision and Pattern Recognition, Seattle, WA, USA, June 2020.

Existing Methods

Multi-scale temporal relation modeling &
Local spatial feature extraction



Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., & Feng, B. (2021). Context-Sensitive Temporal Feature Learning for Gait Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12909–12918).

Some common ground:

A combination of frame-level (or even part-level) feature extraction and temporal information aggregation

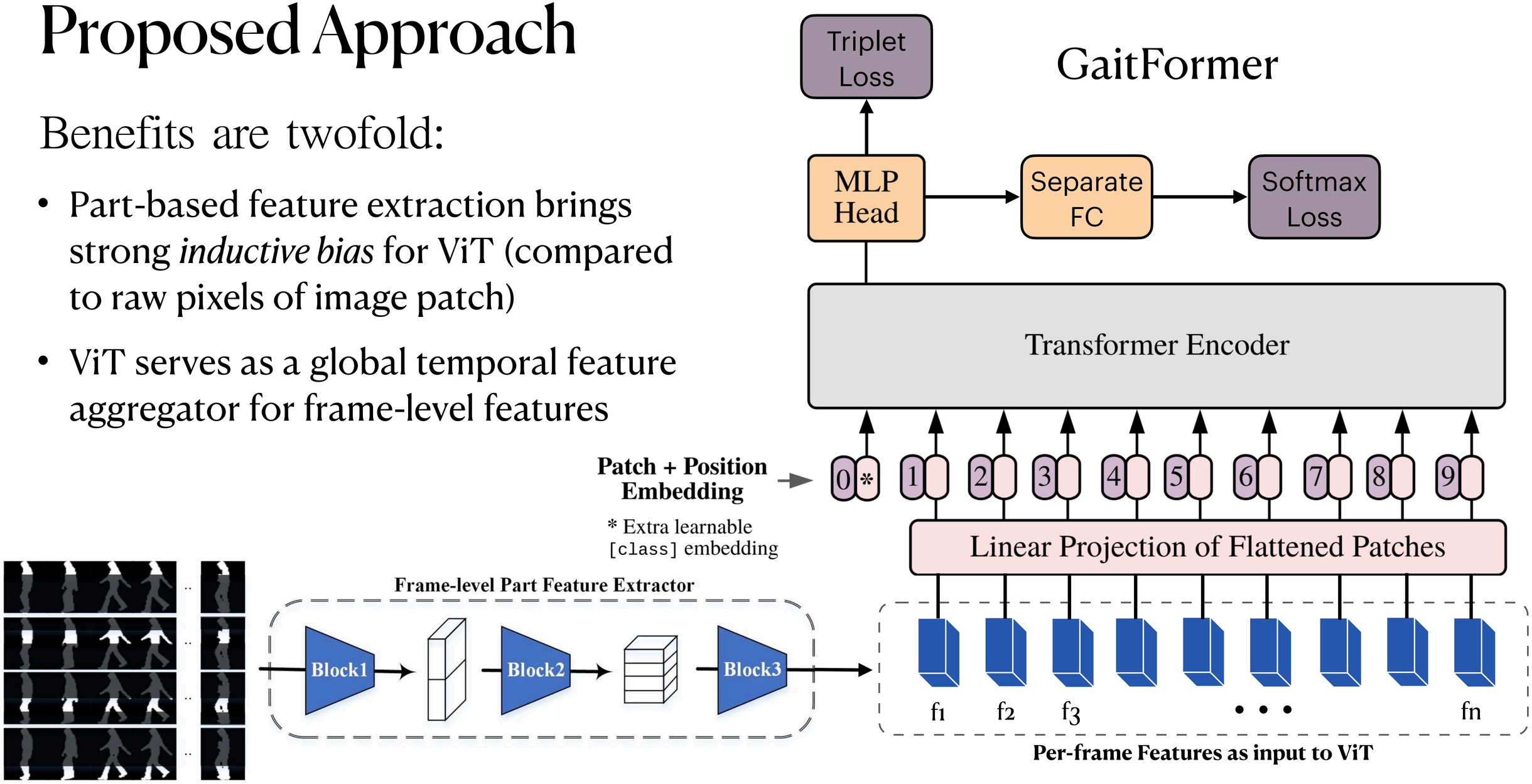
→ Combining both local and global feature is effective for Gait Recognition

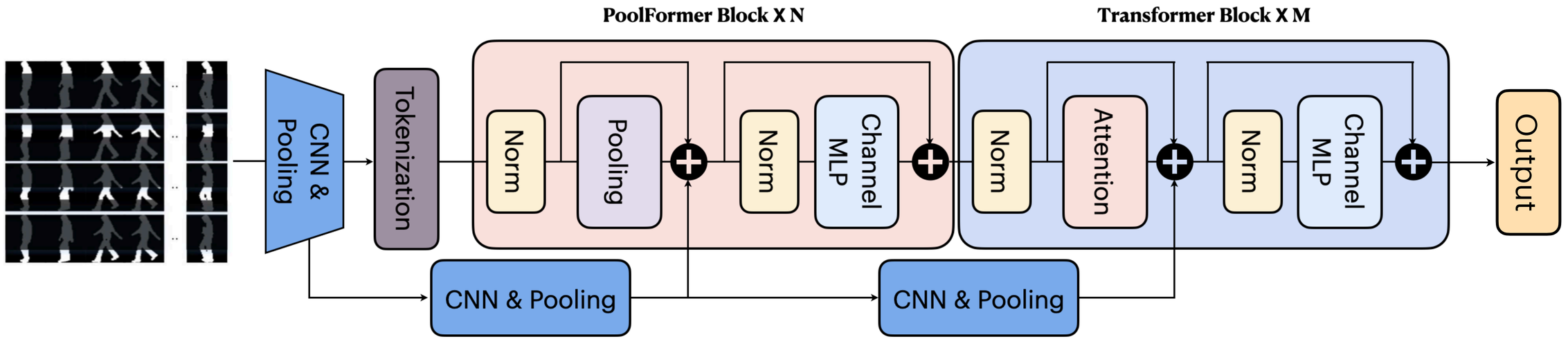
Proposed Approach

Proposed Approach

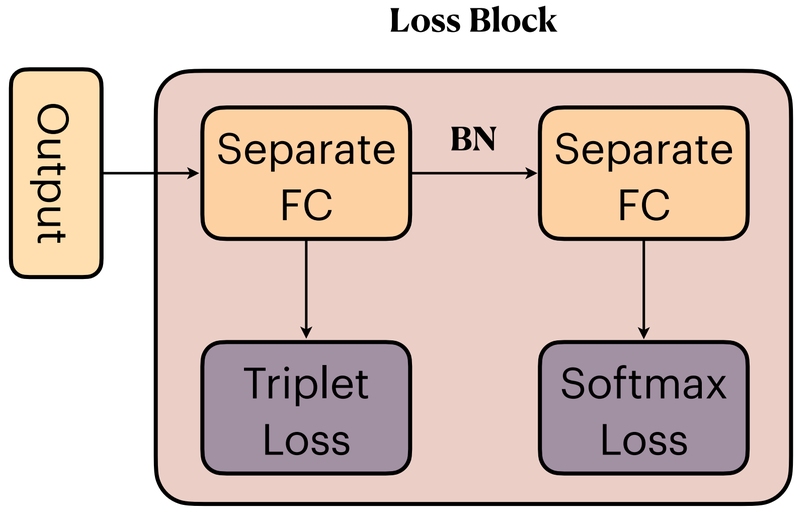
Benefits are twofold:

- Part-based feature extraction brings strong *inductive bias* for ViT (compared to raw pixels of image patch)
- ViT serves as a global temporal feature aggregator for frame-level features





GaitFormerV2



Question:
Is Vision Transformer *all you need* for Gait Recognition?

No! A huge overfitting was observed in initial experiments regardless of model size. It is still hard to train!

Conclusion & Future Works

Conclusion

- Though a combination of part-based feature extraction with vision transformer is expected to be effective, it suffers from overfitting for the present, maybe due to the weak *optimizability* of ViT.

I may next ...

- Use intensive data augmentations/regularizations to improve the *optimizability* of ViT
- Rethink the inductive bias for ViT (are convolution layers in the early stage needed?)
- Explicitly embed order information in tokens instead of in an implicit manner (i.e. use learnable parameter)
- Try to further fuse local and global information by short cut (like ResNet) or direct concatenation

Thanks for listening!