LEARNING TO COUNT EVERYTHING: TRANSFORMER-BASED TRACKERS ARE STRONG BASELINES FOR CLASS AGNOSTIC COUNTING

Anonymous authors
Paper under double-blind review

1 APPENDIX

In the Appendix, we show the introduction and some details of TransT developed by Chen et al. (2021) and Mixformer developed by Cui et al. (2022) as follows for the self-explanatory purpose.

1.1 TRANST

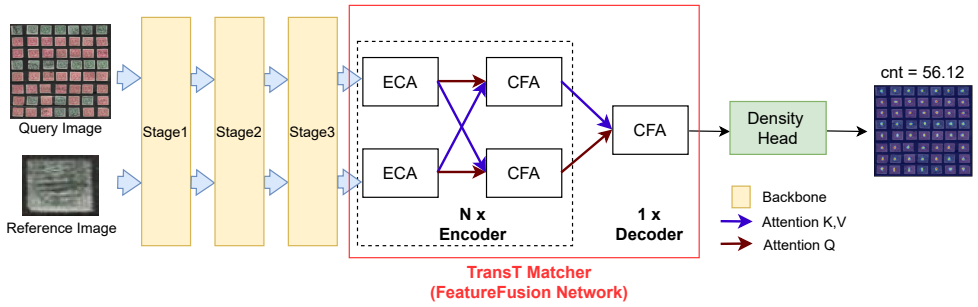

Figure 1: **TransTCAC Model Architecture**. We connect TransT's ResNet50's backbone and matcher onto CAC density head.

The architecture of TransT is illustrated in Figure 1. Given reference and query images, third-layer features (stage3) generated by the ResNet-50 backbone are outputted. The feature fusion network (matcher) is then responsible for localizing target objects within the query image. The prediction head then utilizes the matched features and yields the bounding box predictions. Regarding our needs to enhance matching, we focus on the feature-fusion network and the strength of its sub-components, ego-context augment (ECA) and cross-feature augment (CFA) modules. From a high-level perspective, the feature-fusion network is composed of an encoder for feature refinement and decoder for matching. The encoder individually strengthens feature maps on query and reference branches through ECA (self-attention). Proceedingly, the CFA (cross-attention) extracts foreground pixels on the two branches, exchanging pixel information across branches through dot-product attention.

ECA simulates self-attention and strengthening the query feature maps and reference feature maps individually. On the other hand, CFA, a modified version of cross-attention, selects useful features based on global interactions across query and reference features. Specifically, the CFA module within the encoder's reference branch strengthens regions within reference features having larger similarity with the query, therefore amplifying the foreground reference features while suppressing the noises and background context unnecessary for matching. Likewise, the top CFA module in the encoder's query branch highlights potential pixel candidates of the reference object within the query image. The CFA block in the encoder utilizes cross attention to highlight foreground information for both reference and query features. Given that (1) the feature fusion encoder can effectively determine regions of interests within the feature maps and (2) the feature fusion decoder can perform accurate matching through cross attention, we decide to integrate it onto the CAC matching framework.

1.2 MIXFORMER

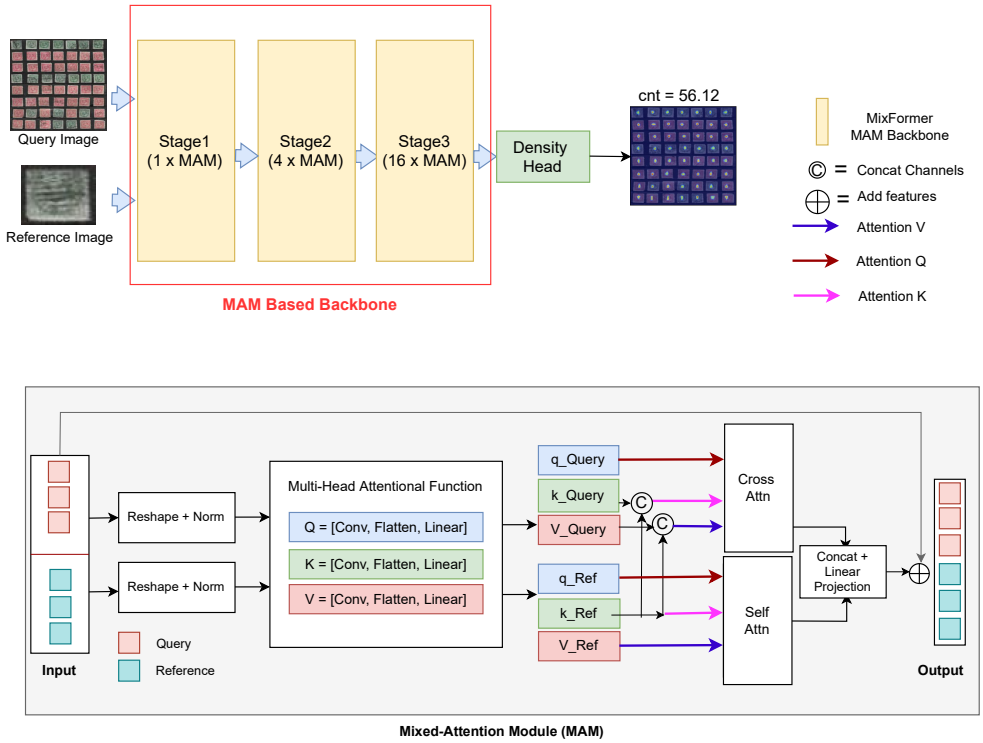

Figure 2: **MixFormerCAC Architecture**. We connect MixFormer's MAM backbone and replace the original prediction head with the CAC density head.

The architecture of MixFormer is illustrated in Figure 2. MixFormer uses CvT proposed by Wu et al. (2021) as backbone, where the input query and references are first mapped into overlapped patch embedding with Convolutional Token Embedding. The patch embeddings are then flattened, concatenated, and inputted into a target-search MAM (Mixed-Attention Module) to perform both feature extraction and information incorporation. Finally, the split and reshaped search tokens are fed into the localization head to gets bounding box coordinates. With regards to feature extraction and matching, asymmetric attention applies self-attention on reference branch and preserves distinctive reference features, while applying cross-attention simultaneously to fuse interactions between query and references and therefore enhance localization and similarity measures.

REFERENCES

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8126–8135, 2021.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13608–13618, 2022.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, and Lei Yuan, Lu; Zhang. Cvt: Introducing convolutions to vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, 2021.