Multiscale Vision Transformers meet Bipartite Matching for efficient single-stage Action Localization
Supplementary Material

Ioanna Ntinou[†,1] Enrique Sanchez[†,2] Georgios Tzimiropoulos[1,2]
[1]Queen Mary University London, UK [2]Samsung AI Center Cambridge, UK
[†]Equal contribution
i.ntinou@qmul.ac.uk kike.sanc@gmail.com g.tzimiropoulos@qmul.ac.uk

1. Architecture details

Herein, we provide the architecture details of our approach when using both the MViTv2-S [6] and ViT-B [8] as backbones. Please refer to [6], [8] and to https://github.com/IoannaNti/BMViT for full implementation details.

stage	operators	output sizes
data	stride 4×1×1	16×256×256
cube$_1$	3×7×7, 96 stride 2×4×4	96×8×64×64
scale$_2$	[MHPA(96) MLP(384)] ×1	96×8×64×64
scale$_3$	[MHPA(192) MLP(768)] ×2	192×8×32×32
scale$_4$	[MHPA(384) MLP(1536)] ×11	384×8×16×16
scale$_5$	[MHPA(768) MLP(3072)] ×2	768×8×16×16
proj	1×1×1, 512	512×8×256
head	MLP(512) MLP(512) MLP(512)	512 × 4 512 × 2 512 × C

(a) Our pipeline with MViTv2-S as bockobone.

stage	operators	output sizes
data	stride 4×1×1	16×288×288
cube$_1$	3×16×16, 768 stride 2×16×16	768×8×18×18
cube$_2$	[MHSA(768) MLP(3072)] ×12	768×8×18×18
head	MLP(768) MLP(768) MHCA(768) MLP(3072) MLP(768)	768 × 4 768 × 2 768 × C

(b) Our pipeline with ViT-B-S as backbone.

Figure 1. 1a The network architecture resembles that of MViTv2-S [6], with the pooling layer after scale$_4$ removed. The output features are projected to 512 dimensions and forwarded to three parallel heads that predict for each token the bounding box coordinates, the probability of the bounding box being an actor, and the class predictions. 1b The network architecture resembles that of ViT-B [6]. The output tokens corresponding to $t = \lfloor T/2 \rfloor$ are forwarded to two parallel heads that predict for each token the bounding box coordinates and the probability of the bounding box being an actor. For the class prediction, we apply cross-attention between all output tokens of shape $8 \times 18 \times 18$ and the ones corresponding to the central frame. The attended tokens are then passed through an MLP for class predictions.

2. Per-class analysis

In Fig. 2 we present the performance per-category of our single-stage model built on MViTv2-S [6] and the corresponding two-stage approach of the same MViT on AVAv2.2 [4]. To compute the per-class accuracy, we re-trained the backbone using the same settings as in [6], obtaining 27.05 mAP, which is indeed $+0.2$ w.r.t. the reported results. Our approach demonstrates improvements in 44 out of the 60 categories, notably increasing in categories like 'drive (e.g., a car, a truck)' with an impressive $+23.1$ mAP increase and 'turn (e.g., a screwdriver)' with a significant $+12.8$ mAP boost. Intriguingly, the performance trends

Method	Pretraining	mAP	GFLOPs	Res.	Backbone	End-to-end
VideoMAE [8]	K400, VideoMAE	31.8	180+246	224	ViT-B	✗
VideoMAE [8]	K400, VideoMAE	37.0	597+246	224	ViT-L	✗
UMT [5]	K400, UMT	32.7	180+246	224	ViT-B	✗
UMT [5]	K400, UMT	39.0	596+246	224	ViT-L	✗
Hiera-L [7]	K400, MAE	**39.8**	413+246	224	Hiera-L	✗
TubeR [10]	K400, IG65M [2]	29.2	97	256	CSN-50	✓
TubeR [10]	K400, IG65M [2]	33.4	138	256	CSN-152	✓
EVAD [1]	K400, VideoMAE	32.3	243	288	ViT-B	✓
STMixer [9]	K400, VideoMAE	32.6	N/A	256	ViT-B	✓
Ours	K400, MAE	38.5	650	256	Hiera-L	✓

Table 1. Comparison w.r.t. state-of-the-art (mean Average Precision; mAP ↑) on AVA v2.2 [3]. "Res." denotes frame resolution.

across categories remain consistent between our method and the two-stage MViT, suggesting the feasibility of employing the same representation for both actor localization and action detection tasks.

3. Additional results

Larger backbones As mentioned in the main document, we trained our method using a Hiera-L backbone. The results of our method against large-scale state-of-the-art methods are shown in Tab. 1. We observe that while our method falls short w.r.t. the two-stage counterpart, our simple, single-stage approach delivers competitive results.

Additional bounding boxes In Eq. (3) in the main document, the training objective considers the bounding box error w.r.t. only the annotated bounding boxes. We want to note that because we distinguish between an actor and a person, the bounding box loss could be also backpropagated w.r.t. bounding boxes corresponding to no-actors, if available. This could improve the precision of the detections and as such the overall performance. We conducted an experiment to study the impact of such approach, observing no effect (neither positive nor negative), in the results. We attribute this to the fact that the number of bounding boxes that correspond to a no-actor on AVA 2.2. is rather small compared to those of the annotated actors.

4. Visual analysis

In Fig. 3 we show three visual examples complementing those presented in Fig. 3 in the main paper. In these examples, we also show the attention maps for each of the tokens that are chosen to represent an actor with $p(\alpha) > \theta$. We can see that the attention maps show how each token is indeed tracking the actor for which they carry the corresponding information, illustrating how our approach can enable the visual tokens to carry bounding box information regarding the central frame, as well as the class information that requires spatio-temporal reasoning.

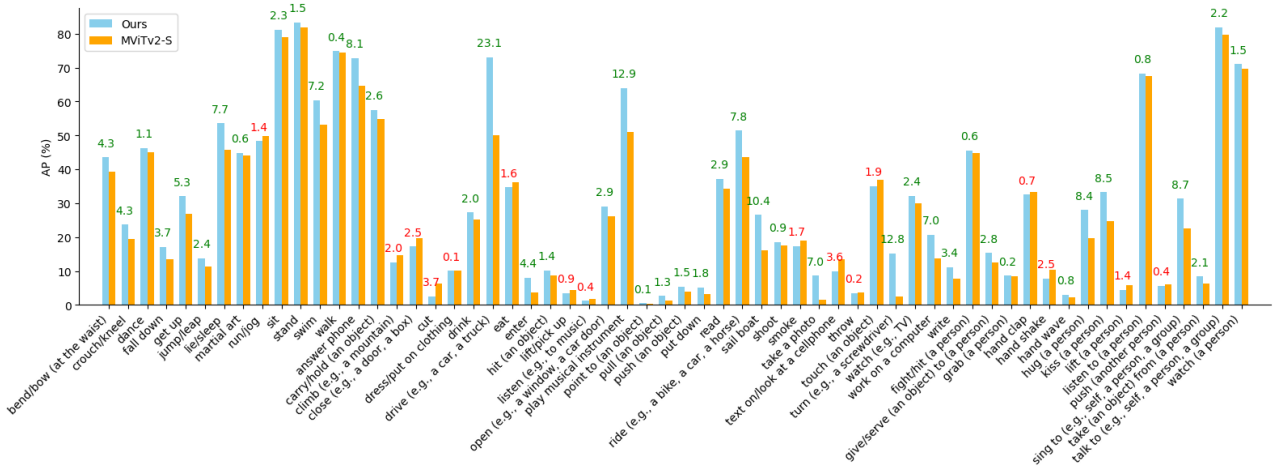

Figure 2. Per-category AP for **Our** single stage action detection method (30.0 mAP) and **MViTv2-S** (27.0 mAP) on AVA v.2. On top of the bar there is the difference per-class where categories with increased accuracy are marked in **green** and those decreased with our method in **red**.

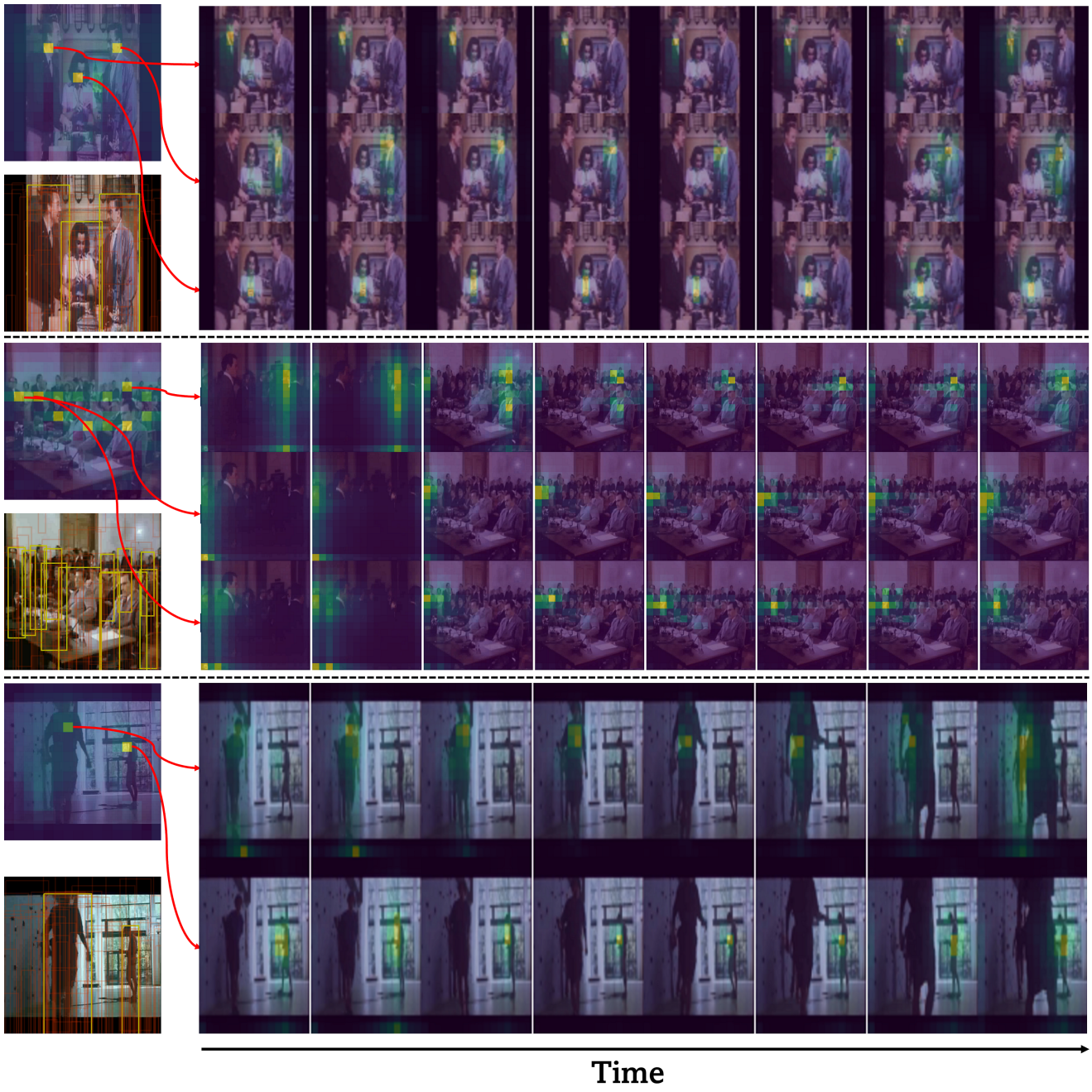

Figure 3. Qualitative analysis (better seen in color and zoomed in). We provide three qualitative examples from three corresponding validation videos from AVA2.2. On the **left top** image we represent the confidence scores $p(\alpha)$ for the actor-no actor prediction, for each of the 16×16 output tokens corresponding to one of the central frames. Those with high confidence $p(\alpha) > \theta$ are selected as positive examples, and their corresponding bounding boxes and class predictions will then form the final outputs. The images in the **left bottom** are the bounding boxes predicted by each of the same 16×16 output tokens, with those in yellow corresponding to the positive tokens (i.e. to the final output bounding boxes). On the **right** we represent the last layer's attention maps corresponding to each of the selected tokens in the left, i.e. their attention scores w.r.t. the whole $8 \times 16 \times 16$ spatio-temporal tokens. We observe that the confident tokens not only attend to the central information to produce the bounding box, but also track the corresponding actor across the video to estimate the corresponding actions. In the second example, we only represent three actors for the sake of clarity. We observe that even with a change of scene, the attention maps can properly track each actor's information. In the last example, the self-attention maps show how they can track each actor's despite the self-occlusion. These examples show that our method can track the actor's information and regress the bounding boxes with a Vision Transformer that assigns each vision token a different output, which are assigned to the ground-truth set through bipartite matching.

References

[1] Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *ICCV*, 2023. 2

[2] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Heng Wang Xueting Yan, and Dhruv Mahajan. Large-scale weaklysupervised pre-training for video action recognition. In *CVPR*, 2019. 2

[3] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[4] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv*, 2020. 1

[5] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, 2023. 2

[6] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 1

[7] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *arXiv*, 2023. 2

[8] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1, 2

[9] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *CVPR*, 2023. 2

[10] Jiaojiao Zhao, Xinyu Li, Chunhui Liu, Shuai Bing, Hao Chen, Cees GM Snoek, and Joseph Tighe. Tuber: Tube-transformer for action detection. *arXiv*, 2021. 2