Supplementary file:
Few-Shot Transformation of Common Actions into Time and Space

Pengwan Yang, Pascal Mettes, Cees G. M. Snoek
University of Amsterdam

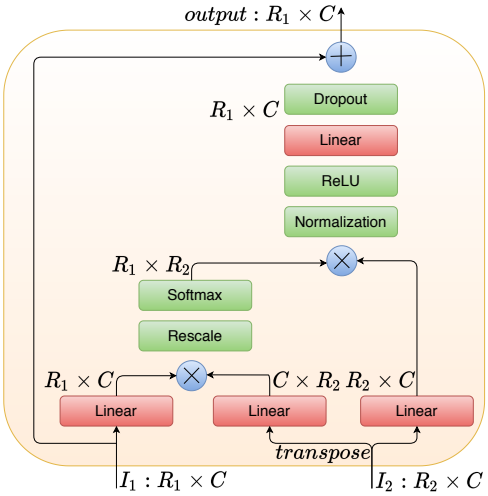

Figure 1: **Overview of common attention block.** I_1, I_2 denote the inputs of our common attention block. $\otimes$ denotes matrix multiplication and $\oplus$ is element-wise sum. The main idea of the common attention block is to align the feature I_2 to the feature I_1.

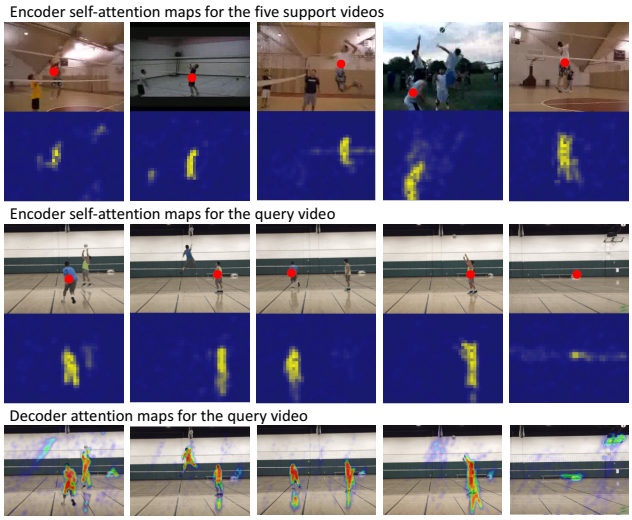

Figure 2: **Attention maps.** We show support and query encoder self-attention maps (top and middle), as well as query decoder attention maps (bottom). • denotes reference point for self-attention map. The encoder makes individual actions stand out in the support and query videos. The decoder highlights the common actions in the query video.

1. Detailed architecture

Overview of the common attention block. The structure of the common attention block is illustrated in Figure 1. The main idea of the common attention block is to align the feature I_2 to the feature I_1. In our model, the common block plays two important roles: i) it aligns each query clip feature with its previous clip features to contain more motion information, ii) it fuses the support feature into the query clip feature based on the joint commonality.

Spatio-temporal positional encoding. In the encoder layers, both support and query branches are associated with corresponding spatio-temporal positions of video features. We generalize the original positional encoding [1] to the 3D case. For all the spatio-temporal coordinates of each embedding, we independently use $\frac{C}{3}$ sine and cosine functions with different frequencies. We then concatenate them to get the final C channel positional encoding.

Visualization. We visualize the attention maps in Figure 2 to better understand our model. The encoder self-attention maps are from the last encoder layer of a trained model. The decoder attention maps are the normalized attention score maps in the common attention block of the decoder. The figure shows that the encoder can make individual actions stand out in the support and query videos, which boosts commonality extraction for the decoder. On the basis of the encoder, the decoder is able to highlight the common actions in the query video.

2. Additional ablations

Benefit of query clip feature alignment. We demonstrate the benefit of query clip feature alignment with the common attention block on the spatio-temporal localization performance on Common-AVA in Figure 3. The neighbor

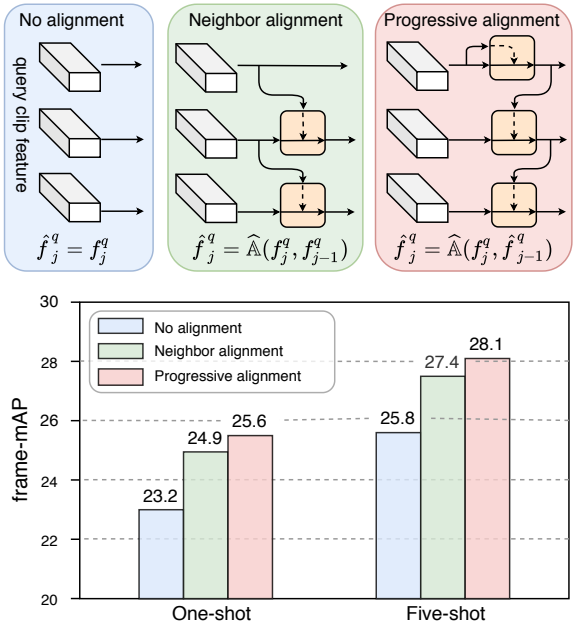

Figure 3: **Benefit of query clip feature alignment.** Propagating the spatio-temporal information from previous clips of the query video into the current query clip by using the common attention block is beneficial to the common localization on common-AVA. Progressively aligning previous query clip features into the current query clip leads to the best results.

Support videos in training	Support videos in evaluation	
All videos are 5 frames	All videos are 5 frames	22.2
	The videos are 5,10,15,20,25 frames	23.8
All videos are 25 frames	All videos are 25 frames	28.1
	The videos are 5,10,15,20,25 frames	25.0
The videos are 5,10,15,20,25 frames	The videos are 5,10,15,20,25 frames	26.1

Table 1: **Effect of variable-length support videos** on five-shot Common-AVA. The results demonstrate our flexibility.

alignment is aligns the current query clip with its *single* previous neighbor query clip feature, while the progressive alignment aligns with *all* previous clip features. So the neighbor alignment lets each query clip contain the spatio-temporal information of its previous neighbor clip. And the progressive alignment propagates long-term motion information of previous clips to the current query clip. The neighbor alignment notably improves the performance and the progressive alignment causes a further performance increase.

Effect of variable-length support videos. We verify our method can handle support videos of varying lengths in Table 1. This is indeed the case, especially when our model is also trained on videos of variable length.

Qualitative results. Some extra qualitative results for common action localization in time and space, and per pixel are shown in Figure 4.

3. Segmentation

The mask-head. Inspired by the extension to segmentation in Carion *et al*. [2], we localize the common action per pixel by simply adding a mask-head upon the decoder outputs, which predicts a binary mask for each of the predicted boxes, see Figure 5. It takes as input the *output embeddings* from the few-shot transformer decoder and computes multi-head attention weights of this embedding over the fused feature of the support and query branches from the encoder, generating attention maps per box in a small resolution. A feature pyramid network architecture [3] is used to increase the resolution and make the final prediction with the supervision of DICE/F-1 loss [5] and Focal loss [4].

Common-A2D. The videos in the dataset have an average length of 136 frames where three to five frames for each video are labeled with dense pixel-level annotations. The selected frames are evenly distributed over a video. There are 2932 videos in the training subset, and 850 videos in the validation and testing subsets. For the training subset, we divide each query video into clips according to the labeled frames, to make each query clip contain one pixel-level annotated frame. Then we sample the query clips to a length of 25 frames. For the validation and testing subsets, we divide each query video into clips of 25 frames long without sampling.

Training details. The mask-head is trained jointly with the whole model for 100 epochs. During inference we first filter out the detection with a confidence below 85% or background label, then compute the per-pixel argmax to determine whether each pixel is foreground.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, 2019. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[5] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 2

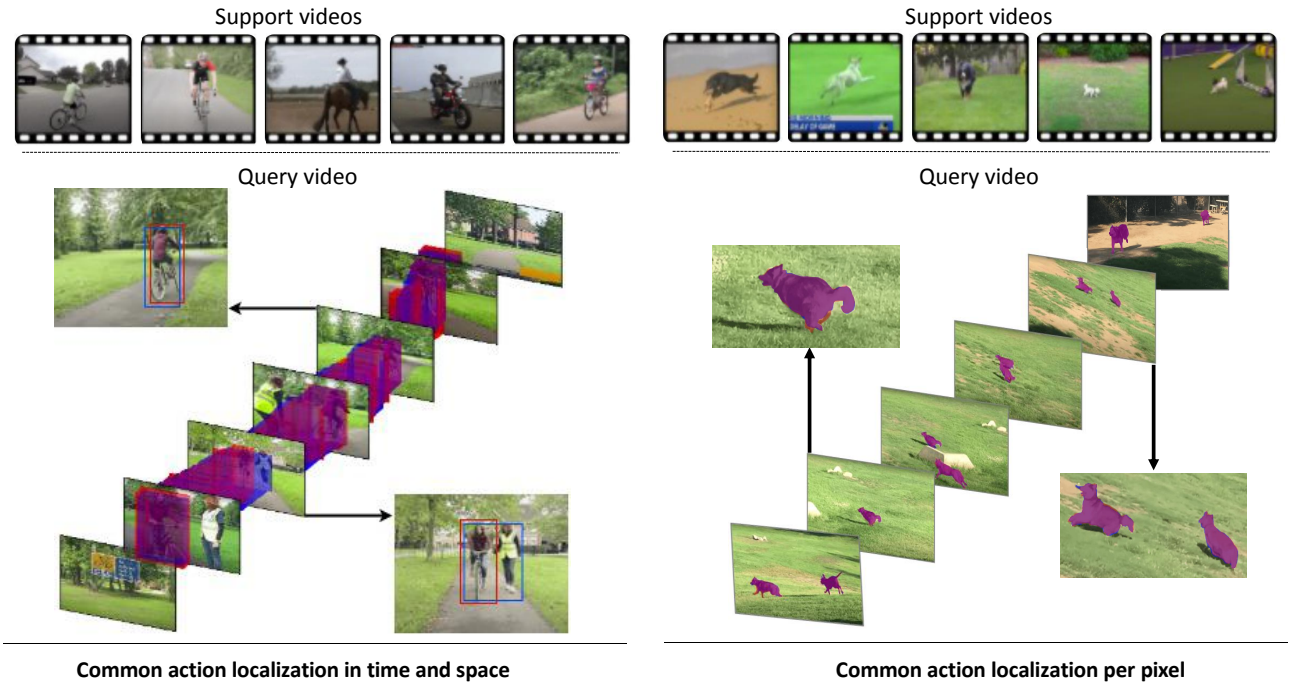

Figure 4: **Qualitative result** under one-shot (blue) and five-shot (red) settings. In the upper part are 2 sets of 5 support videos, where the leftmost video in each set is also used in the one-shot setting. For the left example of common action localization in time and space, with one support video, we can find the common action tube in most clips of the query video, except for the few clips where we wrongly include an extra subject in our prediction. When we use five support videos, our bounding box is refined to exclude the redundant subject. For the right example of common action localization per pixel, with one support video we can localize the common action per pixel, and we present more precise localization with five support videos.

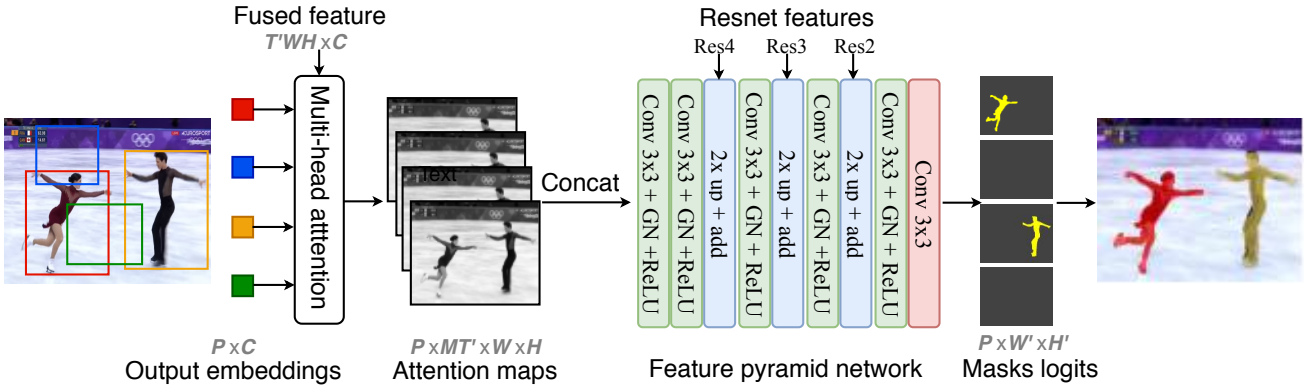

Figure 5: **Structure of the mask-head.** M denotes the head number in the multi head attention, W', H' denote the width and height of the final masks. A binary mask is generated in parallel for each predicted box, then the masks are merged.