Supplementary Material: Mining Inter-Video Proposal Relations for Video Object Detection

Mingfei Han[1,2⋆], Yali Wang[1⋆], Xiaojun Chang[2], and Yu Qiao[1⋆⋆]

[1] Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
{yl.wang, yu.qiao}@siat.ac.cn
[2] Faculty of Information Technology, Monash University, Melbourne, Australia
{hmf282, cxj273}@gmail.com

1 More Implementation Details

To increase receptive fields of ResNet-101, we modify the stride of the first block in the *conv*5 stage from 2 to 1, and all 3×3 convolution layers in this stage is replaced with atrous convolution layers (dilation stride 2).

Specifically, we set 12 sizes of anchor, with 4 scales $\{64^2, 128^2, 256^2, 512^2\}$ and 3 aspect ratios $\{1:2, 1:1, 2:1\}$. During training and testing, we obtain 300 proposals for each frame, via using NMS over 6000 proposals with 0.7 IoU threshold. We use ROI align layer on the feature maps after the *conv*5 stage, and then perform $256 \times 1 \times 1$ convolution layer to obtain the feature vector of each proposal.

Following [2], we use data augmentation in [1] such as random crop, random expand and photometric distortion. Moreover, we resize each input frame to have a shorter side of 600 pixels and longer side of at most 1000 pixels. Due to memory limitations, we adopt a progressive training scheme. We first train detector until intra(2), by using the traditional detection losses of bbox regression and object classification. The learning rate is 0.0005, and drops by a factor of 10 on iteration 110K. Iteration stops at 165K. Then, we freeze convolution layers before *conv*5 and RPN, and finetune the rest modules using the proposed loss, where $\gamma = 1$, and $\lambda = 10$. The learning rate is 0.0016 and drops on iteration 82K. Iteration stops at 110K.

2 More Detection Visualization

We show more detection result of HVR-Net in Fig. 1. We compare two settings, i.e., baseline with only intra-video proposal relation module and our HVR-Net

⋆ Equal contribution.
⋆⋆ Corresponding author.

with both intra-video and inter-video proposal relation modules. With inter-video proposal relations aggregated, our HVR-Net can successfully distinguish all the confusing objects in the video, e.g., the baseline mistakenly recognizes a rabbit in Subplot (a) as rabbit and domestic cat at the same time while our HVR-Net can classify it correctly. The reason is that domestic cat and rabbit share similar appearance and motion characteristics. The inter-video relation could clarify such confusion, while intra-video relation only focus on intra-video similarity in appreance and motion. Hence, it is necessary and important to learn inter-video proposal relations for video object detection.

References

1. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)
2. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: ICCV (2019)

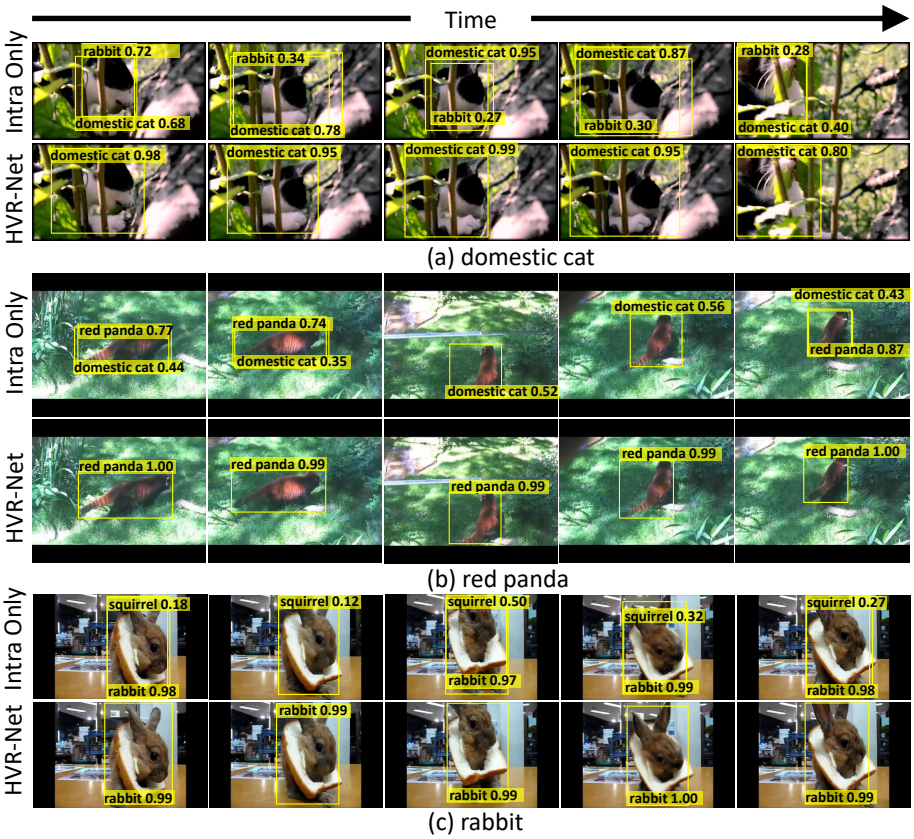

Fig. 1. More Detection Visualization. For each video, the first row shows the baseline with only intra-video proposal relation module. The second row shows HVR-Net with both intra-video and inter-video proposal relation modules. Clearly, our inter-video can effectively guide HVR-Net to tackle object confusion in videos. For example, a domestic cat in Subplot (a) is similar to rabbit in appreance and motion, leanding to high confusion. As a result, the baseline mistakenly recognizes it as a rabbit and domestic cat at the same time, when only using intra-video relation aggregation. By introducing inter-video proposal relation, our HVR-Net successfully distinguish such object confusion in videos. Other subplots also exhibit the similar result, i.e., it is necessary and important to learn inter-video proposal relations to boost video object detection.