Appendix for "PaPr: Training-Free One-Step Patch Pruning with Lightweight ConvNets for Faster Inference"

Tanvir Mahmud[1] Burhaneddin Yaman[2] Chun-Hao Liu[3] Diana Marculescu[1]

[1] University of Texas at Austin
{tanvirmahmud, dianam}@utexas.edu
[2] Bosch Research North America
yaman013@umn.edu
[3] Amazon Prime Video
liuch37@gmail.com

We provide additional qualitative analysis to illustrate the effectiveness of PaPr in practice (Sec. A–Sec. D). Additionally, we provide simple PyTorch pseudo code for PaPr implementation (Sec. E).

A Robustness of PaPr with Various ConvNet Proposals

We study the patch significance map (PSM), and patch masks generated by different ConvNet proposal networks. To make PaPr more computationally efficient and accurate, we need precise mask of discriminative regions irrespective of the size and top-1 accuracy of the proposal network. In Fig. 1, Fig. 2, and Fig. 3, we demonstrate more visualizations of generated PSMs and patch masks with different keeping ratios for various ConvNets. Despite little variations in PSMs for different ConvNets, the top $z\%$ patch mask remains almost identical focusing on the key image patches. With patch keeping ratio of $z = 0.5$, all proposal models mostly keep the key patches representing target objects. With much lower keeping ratio such as $z = 0.3$, few parts of the object in Fig. 2 are masked, however, in Fig. 1 and Fig. 3, most key object patches are visible even with such high masking ratio. This highlights that PaPr can operate with extremely lightweight ConvNet (MobileOne-S0 has 42× smaller FLOPs than ResNet-152) for precise key discriminative patch localization, which makes it particularly suitable for pruning redundant image patches in larger models.

B Robustness of PaPr over CAM Methods

Class activation mapping (CAM) methods mostly focus on highlighting key image patches responsible for the target class prediction to make the decision more interpretable [1, 7, 13, 16, 19]. Such CAM methods have two major limitations preventing their use for patch masking: (1) These methods cannot leverage batch processing for separately tracing the sample activation for each prediction. Moreover, many of these methods rely on gradient modulation [1, 13], or complex activation decomposition [11, 16] which practically makes them infeasible for speeding-up large *off-the-shelf* models. (2) Since these methods heavily rely on activation weights of the final prediction (usually, in the final FC layer), their use is problematic for smaller models with significantly lower top-1 accuracy. Therefore, rather than highlighting the class regions based on final prediction, PaPr attempts to localize the most discriminative patch regions irrespective of their class, thereby making them suitable for ultra-lightweight proposal ConvNets unaffected by their size or final top-1 accuracy.

We provide extensive qualitative comparisons of various CAM methods with PaPr in Fig. 4–Fig. 9. To focus on more challenging samples, we mainly present the results on images, where the proposal ConvNet has significantly lower confidence on the target class while the larger ViT has significantly higher confidence. We denote the final confidence c on the target class in each sample. We use ViT-Base-16 model as the baseline, and lightweight MobileOne-S0 as the proposal model. We analyze whether the application of patch masking with light ConvNet (MobileOne-S0 in our example) has significant impact on the target class prediction of the large model (ViT-Base-16 in our example). We use keeping ratio of $z = 0.4$ to retain top-40% discriminative patches in each method. We highlight several key findings from these qualitative analysis: (1) PaPr can maintain the prediction confidence of large ViTs with significantly smaller amount of patches, whereas other CAM based methods face significant reduction of confidence, mostly in challenging cases. (2) Notably, we observe the increase of prediction confidence in several cases with reduced patches. We hypothesize that such patch masking greatly reduces the complexity of the image by masking the backgrounds, and such simplification of the image increases overall confidence. (3) PaPr performs significantly better particularly in cases where the ConvNet has extremely low confidence, whereas other CAM methods struggle in such scenarios. These demonstrate PaPr's ability to precisely locate the key discriminative patches in challenging scenarios without hurting the large model's performance.

C Qualitative Analysis on Patch Pruning in Videos

In general, video contains large information redundancies, particularly for the video recognition task, that makes such applications computationally burdensome for larger models. However, to locate key discriminative patch regions in videos, spatio-temporal perception of the whole video is required. Interestingly, we can integrate PaPr with light ConvNets for background patch masking with spatio-temporal reasoning to speed-up larger models. In Fig. 10 and Fig. 11, we provide additional visualizations on spatio-temporal patch masking in videos with PaPr. We use lightweight X3d-s [3] model with low patch keeping ratio of $z = 0.3$ for visualization. We highlight the major observations as follows: (1) PaPr can track patches representing the target object across complex backgrounds. (2) In slow moving videos with similar backgrounds, PaPr reduces the data redundancy by suppressing similar frames. Usually, the starting and ending frames are observed with higher priority, while similar intermediate frames are heavily masked. (3) PaPr can precisely isolate few frames representing the main object regions across other redundant frames, thereby requiring holistic understanding of the whole video. These results demonstrate that, PaPr can be very effective in suppressing redundant spatio-temporal patches to significantly reduce computational burden of large *off-the-shelf* models in video recognition.

D Additional Experiments

D.1 More downstream tasks and dynamic spatial pruning methods

The proposed method can be easily extended to downstream tasks like semantic segmentation (SS) and object detection (OD). Initially, we fine-tune a multi-label classifier based on MobileOne-s0 [20] on the target dataset and extract the patch significance map with PaPr. Then, we follow DToP [14], replacing the learnable pruning block with PaPr, where high-confidence tokens exit early and low-confidence tokens go deeper. We fine-tune SegViT [18] for SS and ViTDET [8] for OD using PaPr. Compared to other learnable dynamic spatial pruning methods, reproduced in the same DToP [14] framework, PaPr achieves significantly larger FLOP reduction for similar performance (Tab. I). PaPr can directly speed-up other dynamic methods (*e.g.*, early exiting) as well, by reducing redundant patches before processing, offering a complementary approach to the existing methods (Tab. II).

Tab. I: Comparison on more tasks with ViT-B/16 backbone

Pruning Methods	**Object Det. on COCO**		**Sem. Seg. on ADE20K**	
	box mAP	**GFLOPs**	**mIoU**	**GFLOPs**
None	**51.6**	801	49.6	110
DToP [14]	51.3	659	49.8	86
SAR [5]	51.2	680	49.5	88
LAUDNet [6]	51.3	692	49.4	91
Ours	51.4	**510**	**49.9**	**70**

Tab. II: Integration of PaPr in early exiting methods

Method	**Arch**	**PaPr**	**Acc.**	**Img/s**
Dyn-Perceiver [4]	ConvNeXt-Base1k	✗	**82.3**	385
		✓	82.2	**498**
DVT [17]	ViT-B/16	✗	83.4	552
		✓	83.4	**694**

D.2 Comparison with CAM variants and random pruning at different keeping ratio

Existing CAM variants are mostly limited by batch in-operability for sample-wise optimization, and the intrinsic lower accuracy of ConvNets for using FC layers, which limits their effectiveness for expediting the ViT models. In contrast, PaPr can extract precise masks even for classes where ConvNet confidence is low, and hence it can maintain ViT accuracy while enabling significant speed up (see Fig. 7 in the main paper). We provide additional quantitative analysis with different CAM variants, and random pruning for the patch selection in PaPr at various patch keeping ratio (z) (Tab. III). Random pruning gets the lowest accuracy due to lacking a holistic view and consequently removing essential information. Despite the lower accuracy with smaller keep ratio for random pruning, unlike other CAM variants, random pruning can be processed in batches, which provides additional speed-ups for large-scale processing. In contrast, our method achieves the best performance for all keep ratios while maintaining the batch processing as random pruning for faster operation.

Tab. III: Patch localization methods at various keep ratio (z). We use ViT-B/16 MAE model with MobileOne-S0 patch selector.

Methods	BatchOp.	z=1.0	z=0.7	z=0.6	z=0.5	z=0.4
Random	✓	83.7	79.3	73.5	65.4	55.7
CAM [19]	✗	83.7	80.5	77.9	74.5	71.3
GradCAM [13]	✗	83.7	81.1	78.8	76.7	73.8
ScoreCAM [16]	✗	83.7	80.9	78.5	75.8	73.2
Ours	✓	83.7	**83.3**	**82.9**	**82.4**	**81.4**

D.3 FC layers are bottleneck

In Tab. IV, we examine the effects of FC layers in PaPr with and without the proposed weight suppression. Including FC layers results in significantly lower performance. Additionally, in Fig. 7 from the main paper and Fig. 4–9, we highlight that PaPr can maintain ViT accuracy even when the proposal ConvNet confidence is very low. This show that by suppressing FC layers, we can use ultra-lightweight ConvNets to speed-up large-scale ViTs.

Tab. IV: Effect of FC layers in PaPr at various keep ratio (z)

Models	FC Layer	z=0.7	z=0.6	z=0.5	z=0.4
ViT-B/16	✓	81.2	79.1	77.5	74.8
(& MobileOne-S0)	✗	**83.3**	**82.9**	**82.4**	**81.4**
ViT-L/16	✓	82.4	81.3	80.2	78.6
(& MobileOne-S0)	✗	**85.8**	**85.5**	**85.1**	**84.8**

E PyTorch Implementation

We provide pseudo code implementation of PaPr in PyTorch [12] on class-token based vision transformer [2]. In particular, we apply PaPr to prune redundant patch tokens in ViTs after the initial extraction of patch tokens with the integration of class token and position embedding. Starting from the ViT tokens, and final convolutional feature maps extracted from the proposal ConvNet, the following code snippet can prune redundant patch tokens with target keeping ratio (z). We note that, PaPr can operate with class-token free ViTs [2], hierarchical transformers [9], pure ConvNets [10], and video transformers [15].

```
import torch.nn.functional as F

def apply_papr(x: torch.tensor, f: torch.tensor, z: float) ->
                                torch.tensor:
"""
    x: input ViT tokens of size (batch, N, c)
    f: proposal ConvNet features of size (batch, K, h, w)
    z: keeping ratio for tokens
"""
    b, n, c = x.shape
    h1 = w1 = numpy.sqrt(n-1) # spatial resolution of tokens
    nt = int(n*z) # total remaining tokens after pruning

    # extract discriminative feature map from proposal features
    Fd = f.mean(dim=1) # size (batch, h, w)

    # upsampling F to match patch token spatial resolution in x
    # it generates Patch Significance Map (P)
    P = F.interpolate(Fd, size=(h1, w1), mode="bicubic")
    P = P.view(b, -1) # reshaping for pruning mask extraction

    # extracting indices of the most significant patches
    patch_indices = P.argsort(dim=1, descending=True)[:, :nt]

    patch_indices += 1 # adjusting indices for class tokens

    # preparing class indices for each sample
    class_indices = torch.zeros(b, 1).to(patch_indices.device)

    # Patch mask is obtained combining class and patch indices
    M = torch.cat([class_indices, patch_indices], dim=1)

    # extracting tokens based on patch mask
    x = x.gather(dim=1, index=M.unsqueeze(-1).expand(b, -1, c))

    # pruned x tensor size (batch, nt, c)
    return x
```

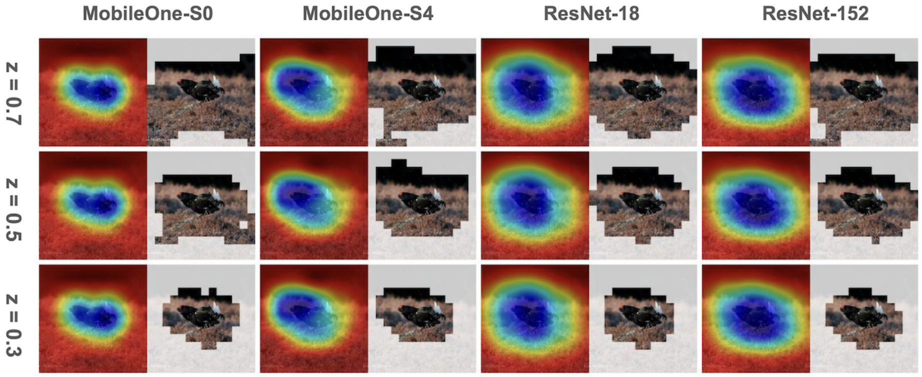

Fig. 1: More visualizations of patch significance map (PSM) and patch masks with various proposal models for different keeping ratio (z).

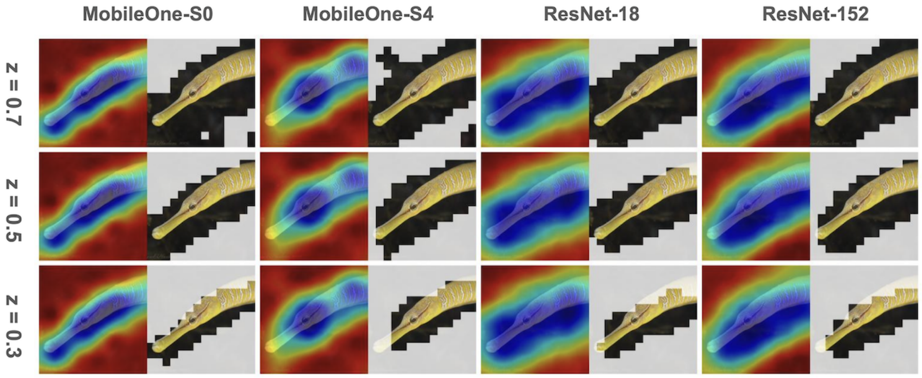

Fig. 2: More visualizations of patch significance map (PSM) and patch masks with various proposal models for different keeping ratio (z).

References

1. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
3. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–213 (2020)
4. Han, Y., Han, D., Liu, Z., Wang, Y., Pan, X., Pu, Y., Deng, C., Feng, J., Song, S., Huang, G.: Dynamic perceiver for efficient visual recognition. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5969–5979 (2023). https://doi.org/10.1109/ICCV51070.2023.00551

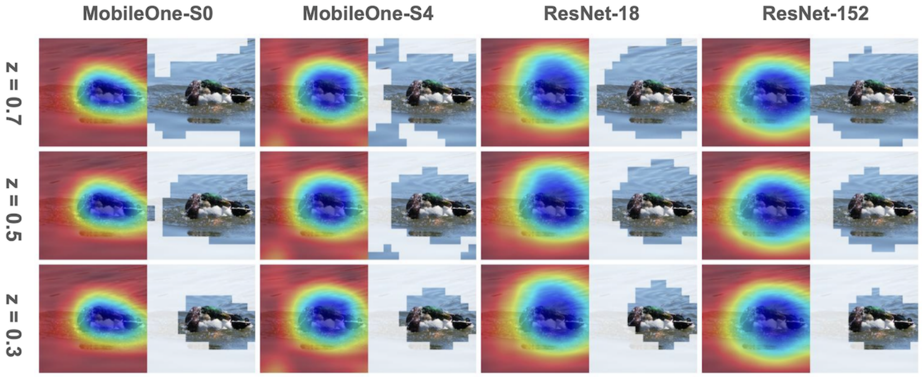

Fig. 3: More visualizations of patch significance map (PSM) and patch masks with various proposal models for different keeping ratio (z).

5. Han, Y., Huang, G., Song, S., Yang, L., Zhang, Y., Jiang, H.: Spatially adaptive feature refinement for efficient inference. IEEE Transactions on Image Processing **30**, 9345–9358 (2021). https://doi.org/10.1109/TIP.2021.3125263
6. Han, Y., Liu, Z., Yuan, Z., Pu, Y., Wang, C., Song, S., Huang, G.: Latency-aware unified dynamic networks for efficient image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–17 (2024). https://doi.org/10.1109/TPAMI.2024.3393530
7. Jung, H., Oh, Y.: Towards better explanations of class activation mapping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1336–1344 (2021)
8. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: Computer Vision âĂŞ ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23âĂŞ27, 2022, Proceedings, Part IX. p. 280âĂŞ296. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20077-9_17, https://doi.org/10.1007/978-3-031-20077-9_17
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
11. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)

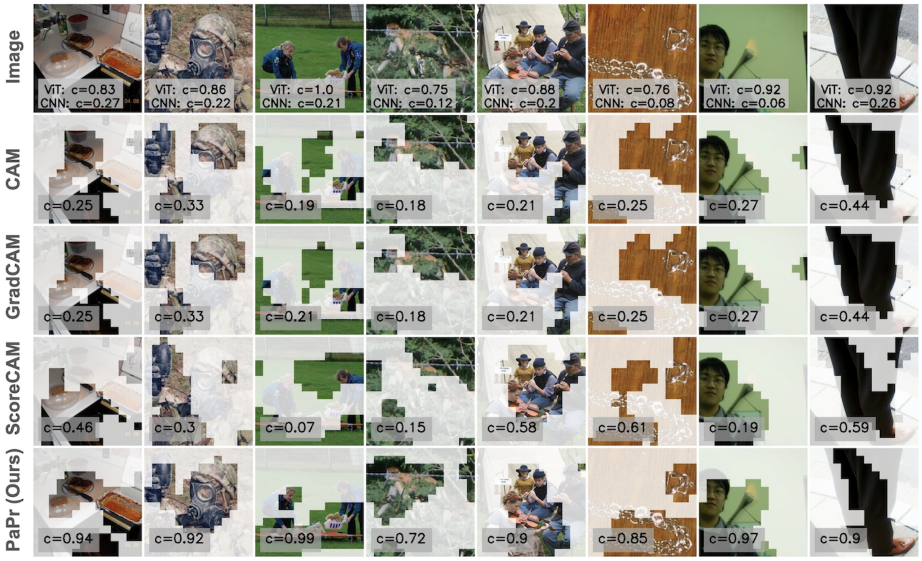

Fig. 4: More visualizations on robustness of PaPr compared to CAM based methods.

14. Tang, Q., Zhang, B., Liu, J., Liu, F., Liu, Y.: Dynamic token pruning in plain vision transformers for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 777–786 (October 2023)
15. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv:2203.12602 [cs.CV] (2022)
16. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
17. Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G.: Not all images are worth 16x16 words: dynamic transformers for efficient image recognition. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2024)
18. Zhang, B., Tian, Z., Tang, Q., Chu, X., Wei, X., Shen, C., Liu, Y.: Segvit: semantic segmentation with plain vision transformers. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2024)
19. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
20. Zhu, K., Wu, J.: Residual attention: A simple but effective method for multi-label recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 184–193 (October 2021)

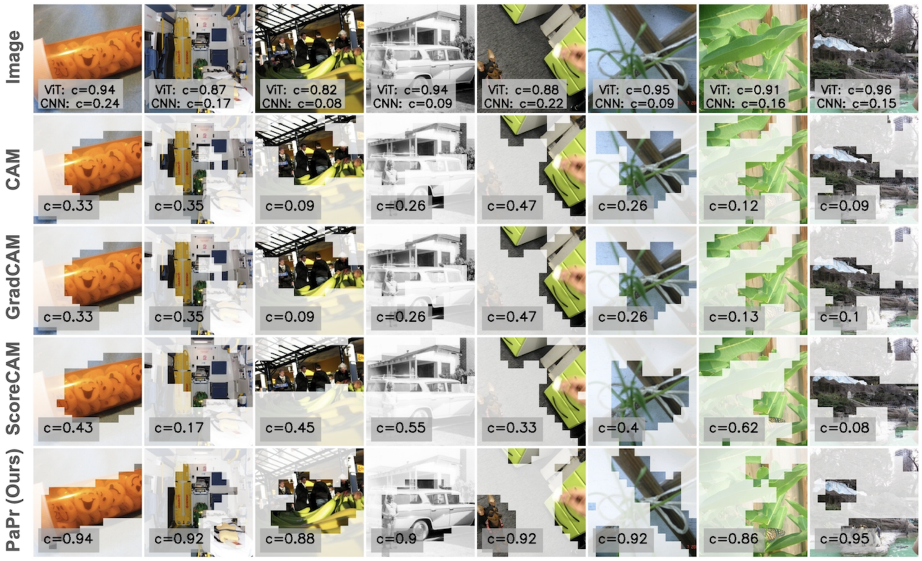

Fig. 5: More visualizations on robustness of PaPr compared to CAM based methods.

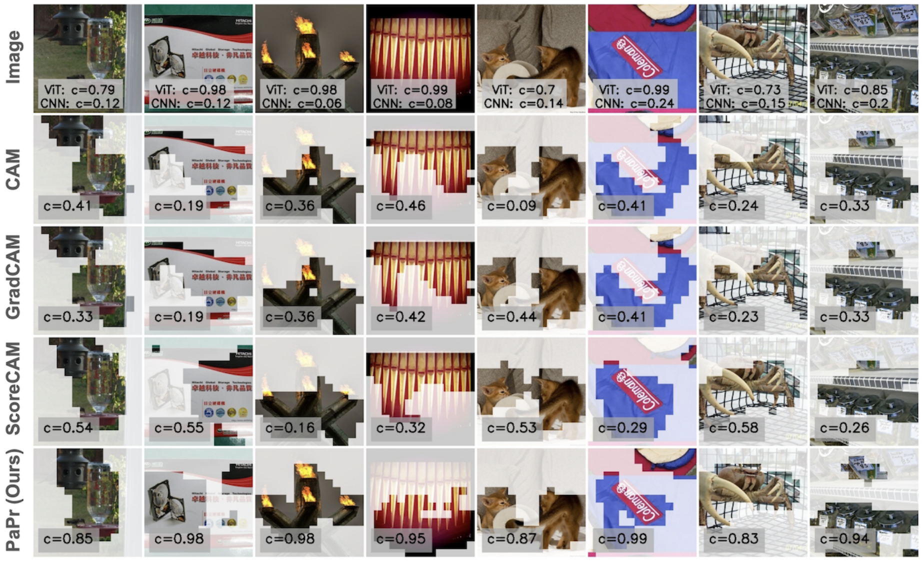

Fig. 6: More visualizations on robustness of PaPr compared to CAM based methods.

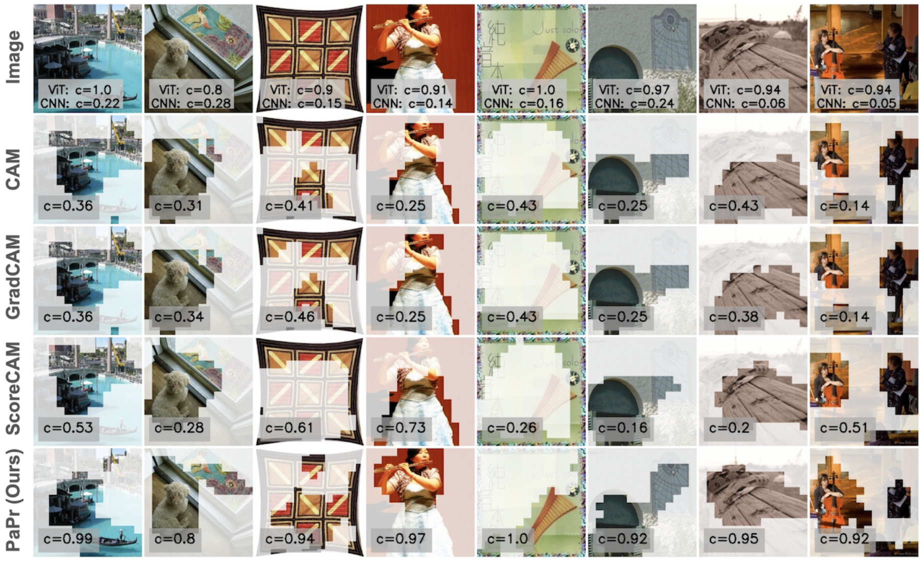

Fig. 7: More visualizations on robustness of PaPr compared to CAM based methods.

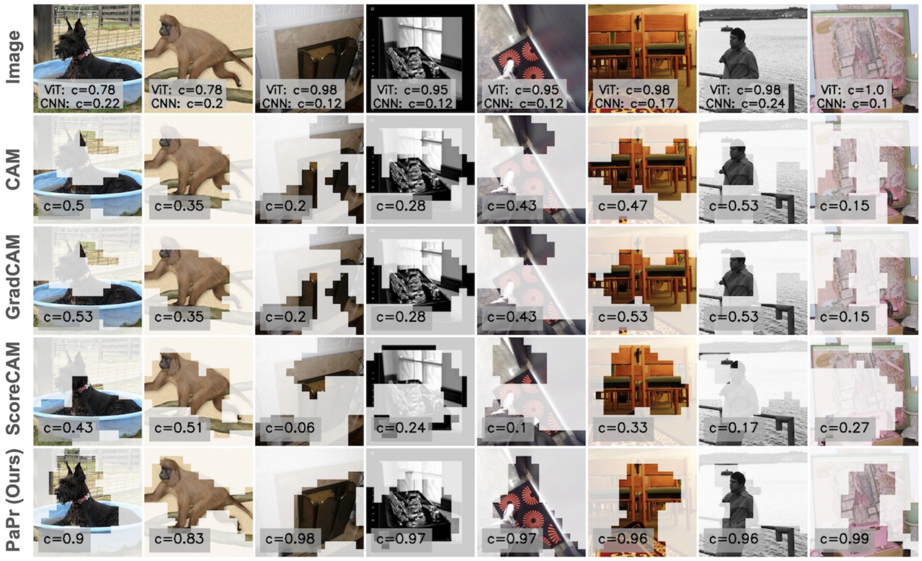

Fig. 8: More visualizations on robustness of PaPr compared to CAM based methods.

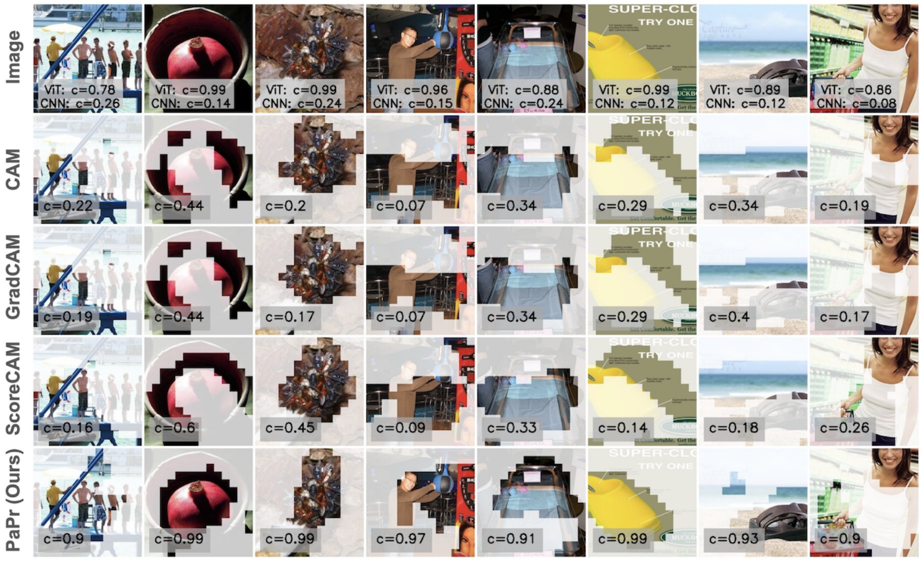

Fig. 9: More visualizations on robustness of PaPr compared to CAM based methods.

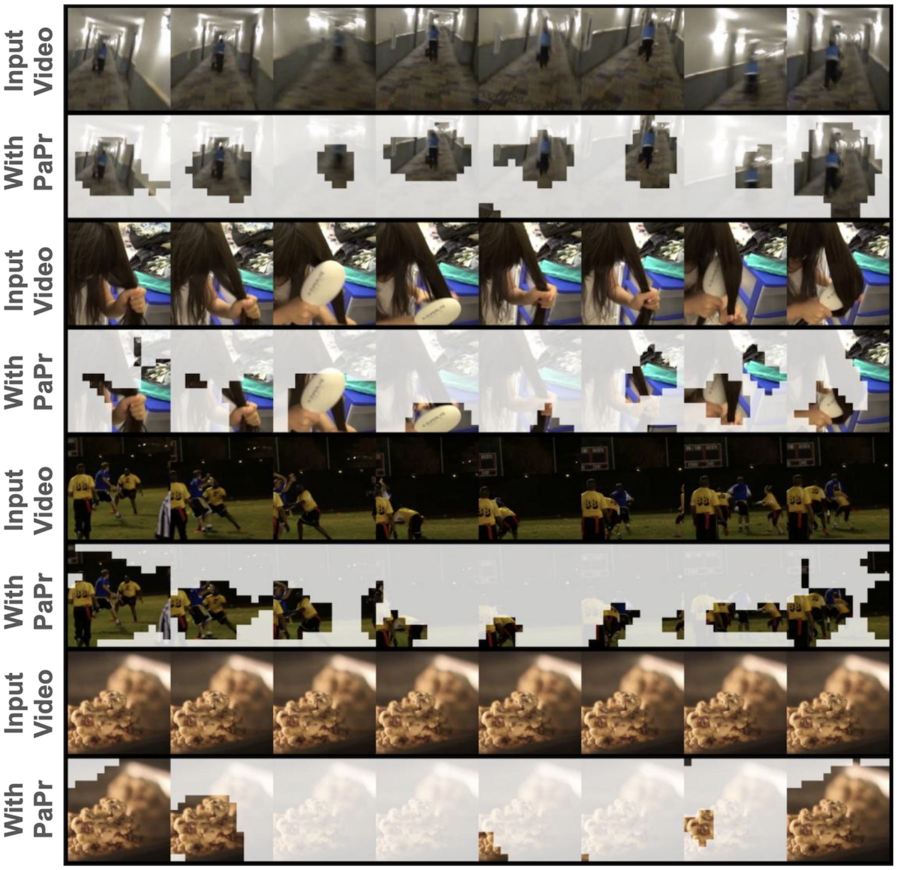

Fig. 10: More visualizations of spatio-temporal patch masking in videos with PaPr for keeping ratio $z = 0.3$. X3d-s [3] based ConvNet is used for visualization.

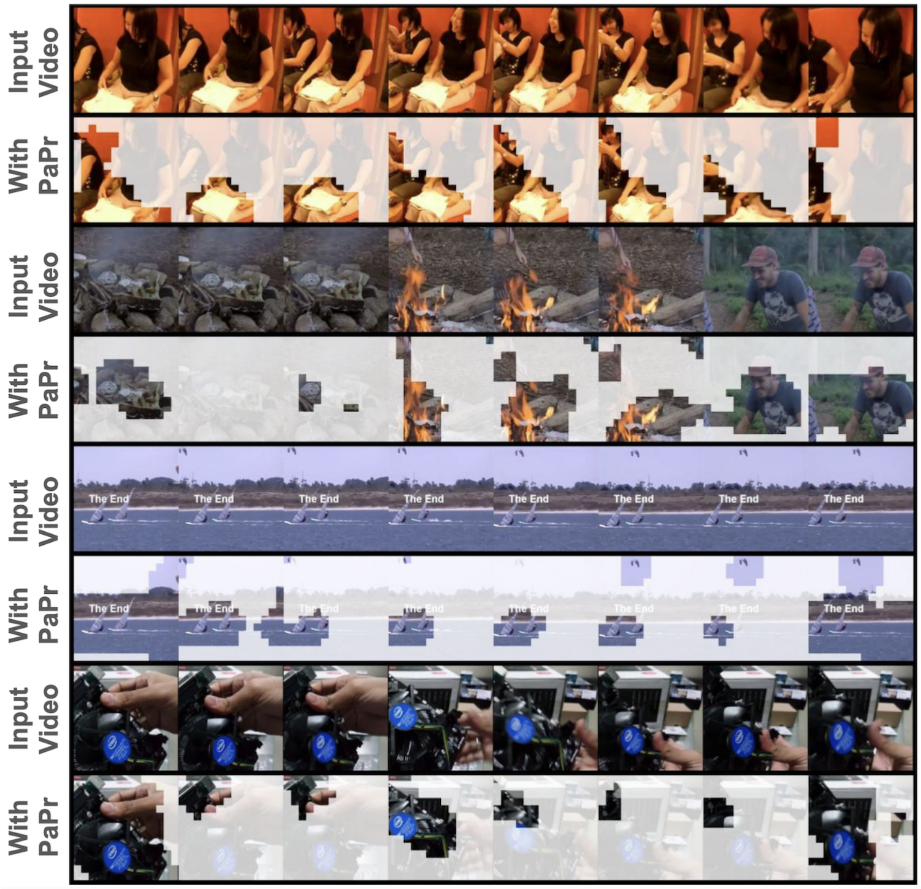

Fig. 11: More visualizations of spatio-temporal patch masking in videos with PaPr for keeping ratio $z = 0.3$. X3d-s [3] based ConvNet is used for visualization.