







AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale – Supplementary Materials

Adam Pardy^{1,2,3}, Michał Wronka², Maciej Wołczyk¹, Kamil Adamczewski¹, Tomasz Trzcinski^{1,4,5}, and Bartosz Zieliński^{1,2}

¹ IDEAS NCBR

{adam.pardyl, maciej.wolczyk, kamil.adamczewski,
tomasz.trzcinski, bartosz.zielinski}@ideas-ncbr.pl

² Jagiellonian University, Faculty of Mathematics and Computer Science
michal.wronka@student.uj.edu.pl

³ Jagiellonian University, Doctoral School of Exact and Natural Sciences

⁴ Warsaw University of Technology

⁵ Tooploox

In this Supplementary Materials, we present additional experiments and visualizations for our AdaGlimpse active visual exploration method.

Table 1: Heuristic baselines: We compared our reinforcement learning-based method against two heuristic exploration policies for the reconstruction task on the ImageNet-1k dataset. Both heuristic methods use the same ElasticViT backbone as our approach and utilize 32^2 glimpses, consisting of four ViT patches. The first method, called BasicElasticAME, samples one low-resolution glimpse of the entire image, followed by small, high-resolution glimpses selected using the AME algorithm. The second method, called QuadTreeAME, also begins with a single low-resolution glimpse. It then iteratively selects the most uncertain patch using AME, sampling a new glimpse at its position to effectively acquire four new higher-resolution patches in place of the selected lower-resolution patch. For comparison, we also provide results for a standard, fixed-resolution AME baseline. We observed that our method outperforms these heuristic baselines.

Method	RMSE	Image res.	Glimpses	Regime	Pixel %
AME	30.3	128×256	8×32^2	simple	25.00
BasicElasticAME	25.95	224×224	12×32^2	adaptive (heuristic)	24.49
QuadTreeAME	23.2	224×224	12×32^2	adaptive (heuristic)	24.49
Ours	14.5	224×224	12×32^2	adaptive	24.49

Table 2: Model and training configuration: In this table we specify the model and training configuration details.

Parameter name	Value
Encoder	
ViT type	base
native patch size	16
transformer embed dim	768
transformer blocks	12
attention heads	12
mlp ratio	4
Decoder	
transformer embed dim	512
transformer blocks	8
attention heads	16
mlp ratio	4
RL agent	
hidden dim	256
action distribution	TanhNormal
sac target entropy value	-3
sac initial alpha value	1
Training	
training epochs	100
backbone pre-training epochs	600
backbone lr	10^{-5}
rl agent lr	$5 * 10^{-4}$
backbone weight decay rate	10^{-4}
rl agent weight decay rate	10^{-2}
lr scheduler type	one cycle
lr warmup epochs	10
minimum lr	10^{-8}
initial random action batches	10000
initial frozen backbone epochs	10
rl loss function	L2
rl batch size	256
backbone batch size	128
replay buffer size	10000

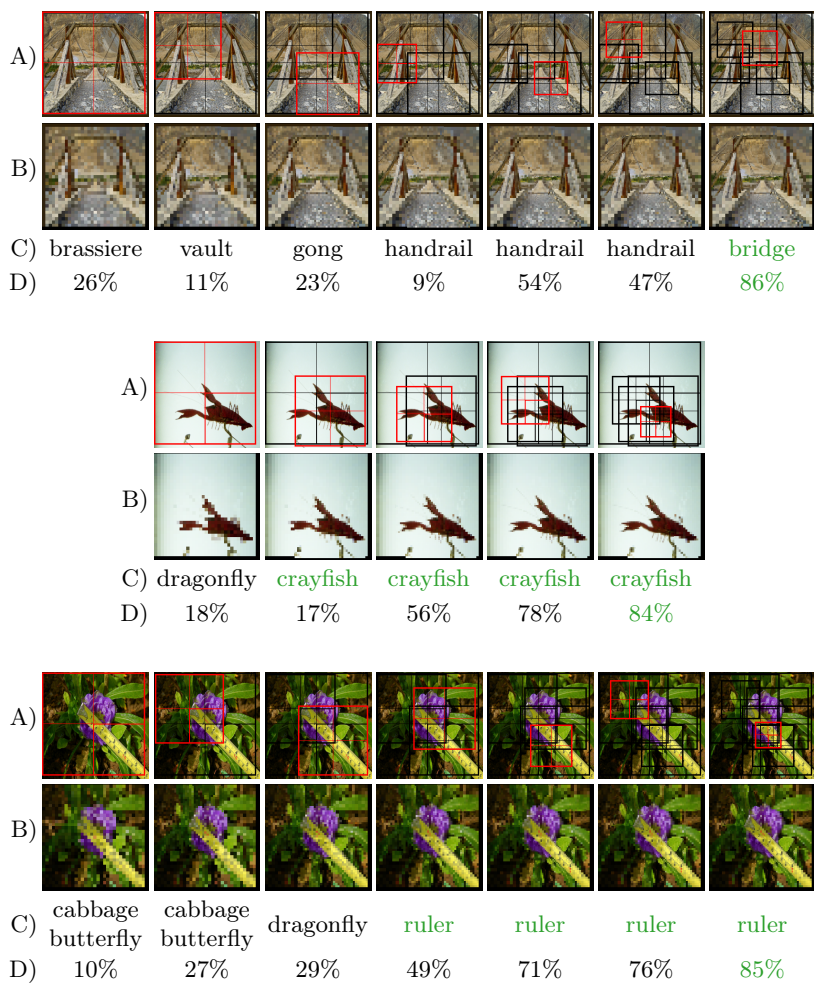


Fig. 1: Image classification on ImageNet-1k step-by-step: AdaGlimpse explores 224×224 images from ImageNet with 32×32 glimpses of variable scale, zooming in on objects of interest and stopping the process after reaching 75% predicted probability. The rows correspond to: A) glimpse locations, B) pixels visible to the model (interpolated from glimpses for preview), C) predicted label, D) prediction probability.

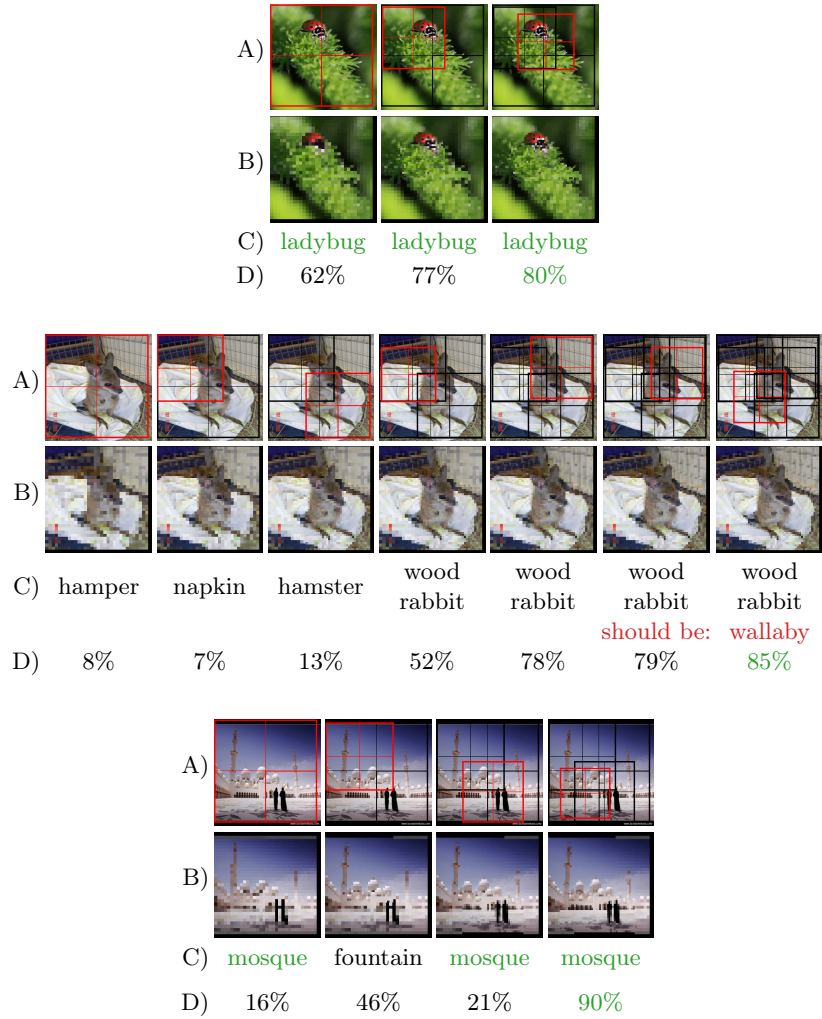


Fig. 2: Image classification on ImageNet-1k step-by-step: AdaGlimpse explores 224×224 images from ImageNet with 32×32 glimpses of variable scale, zooming in on objects of interest and stopping the process after reaching 75% predicted probability. The rows correspond to: A) glimpse locations, B) pixels visible to the model (interpolated from glimpses for preview), C) predicted label, D) prediction probability.

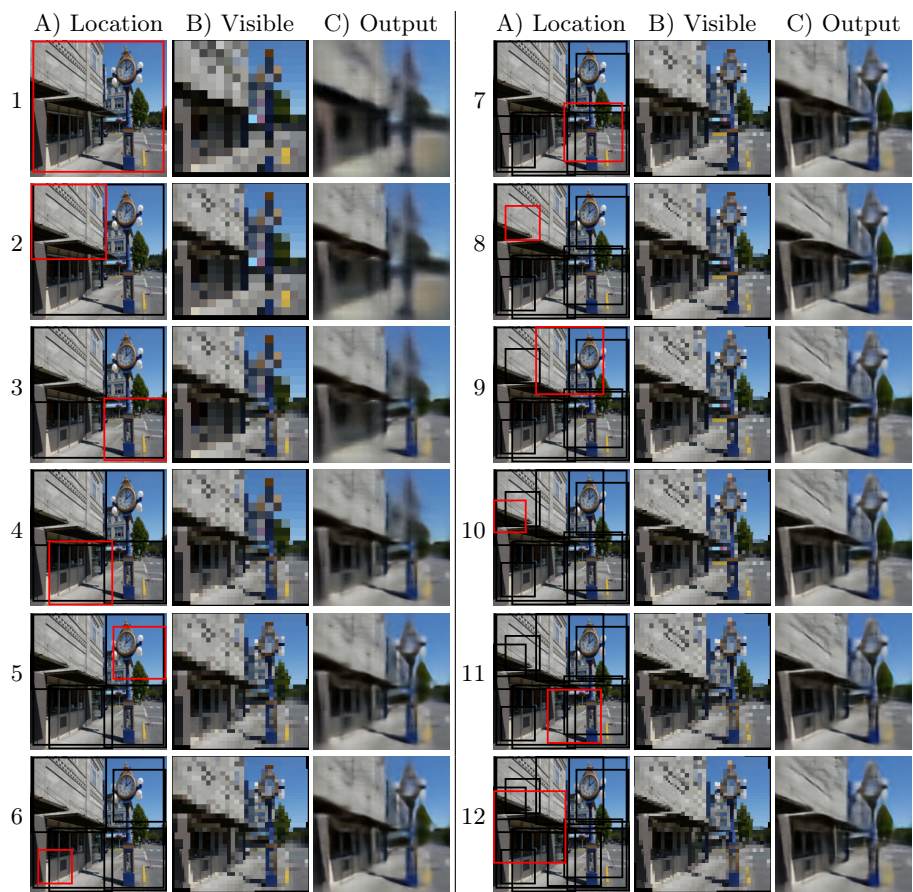


Fig. 3: Image reconstruction on MS COCO step-by-step: AdaGlimpse explores 224×224 images from MS COCO with 8×16^2 glimpses of variable scale, zooming in on objects of interest. Note, that each glimpse consists of a single vision transformer patch. The columns correspond to: A) glimpse locations, B) pixels visible to the model (interpolated from glimpses for preview), C) reconstruction result.



Fig. 4: Image reconstruction on MS COCO step-by-step: AdaGlimpse explores 224×224 images from MS COCO with 8×16^2 glimpses of variable scale, zooming in on objects of interest. Note, that each glimpse consists of a single vision transformer patch. The columns correspond to: A) glimpse locations, B) pixels visible to the model (interpolated from glimpses for preview), C) reconstruction result.

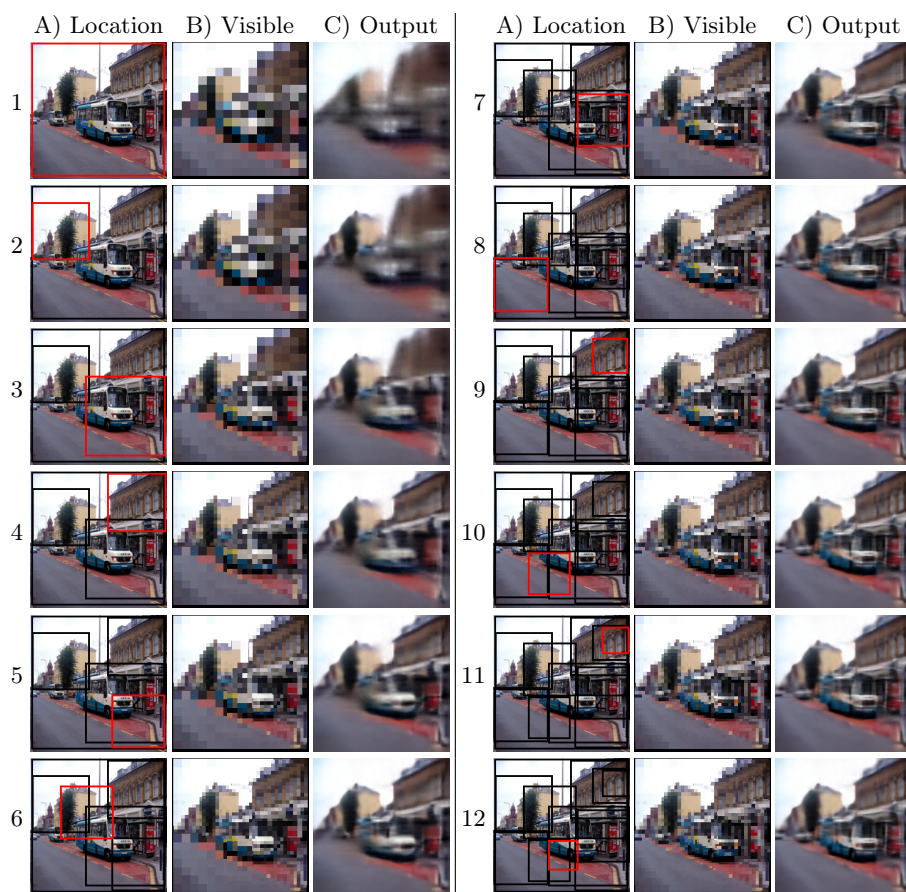


Fig. 5: Image reconstruction on MS COCO step-by-step: AdaGlimpse explores 224×224 images from MS COCO with 8×16^2 glimpses of variable scale, zooming in on objects of interest. Note, that each glimpse consists of a single vision transformer patch. The columns correspond to: A) glimpse locations, B) pixels visible to the model (interpolated from glimpses for preview), C) reconstruction result.

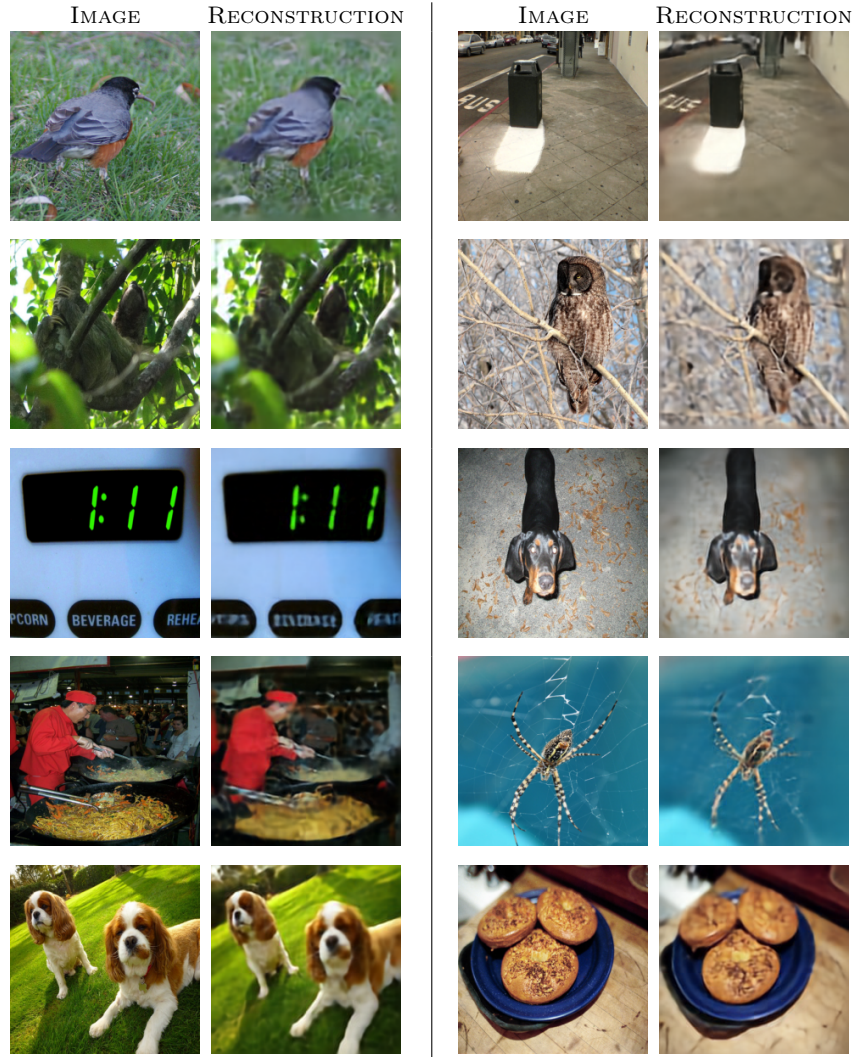


Fig. 6: Image reconstruction on ImageNet-1k: Figure shows the qualitative results of the image reconstruction task. Image size is 224×224 and 12×32^2 adaptive glimpses are used.

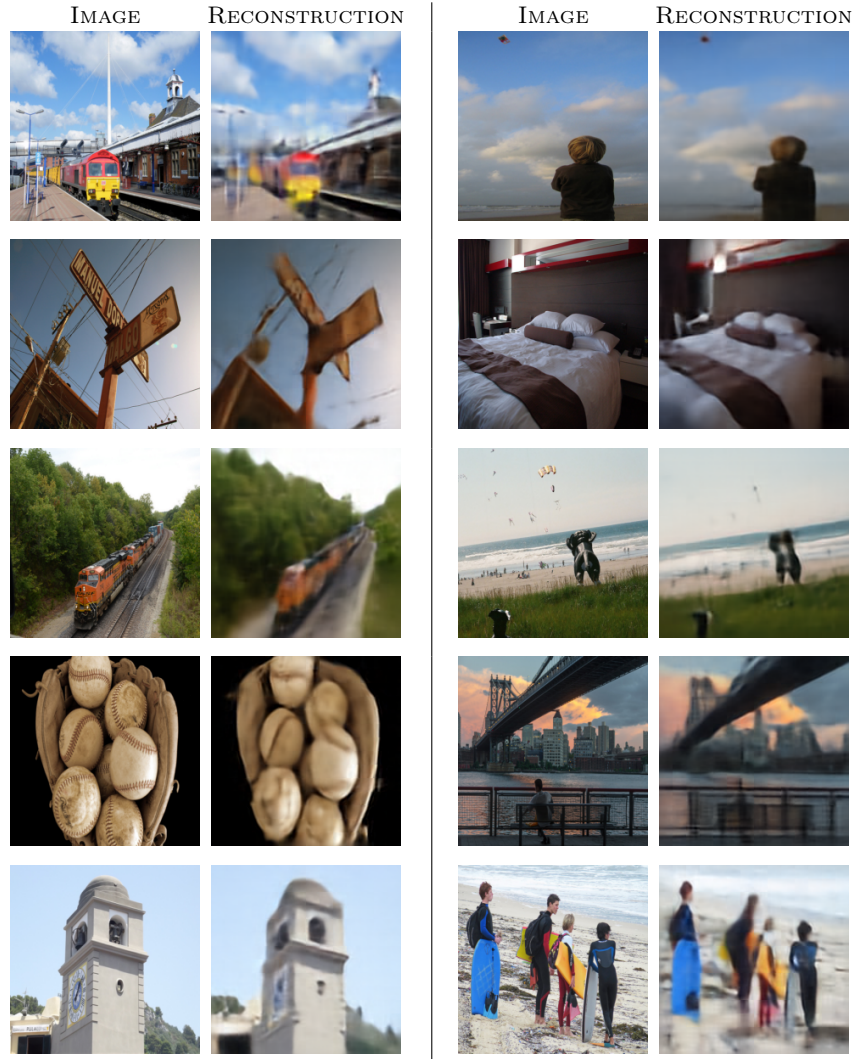


Fig. 7: Image reconstruction on MS COCO: Figure shows the qualitative results of the image reconstruction task. Image size is 224×224 and 12×16^2 adaptive glimpses are used.

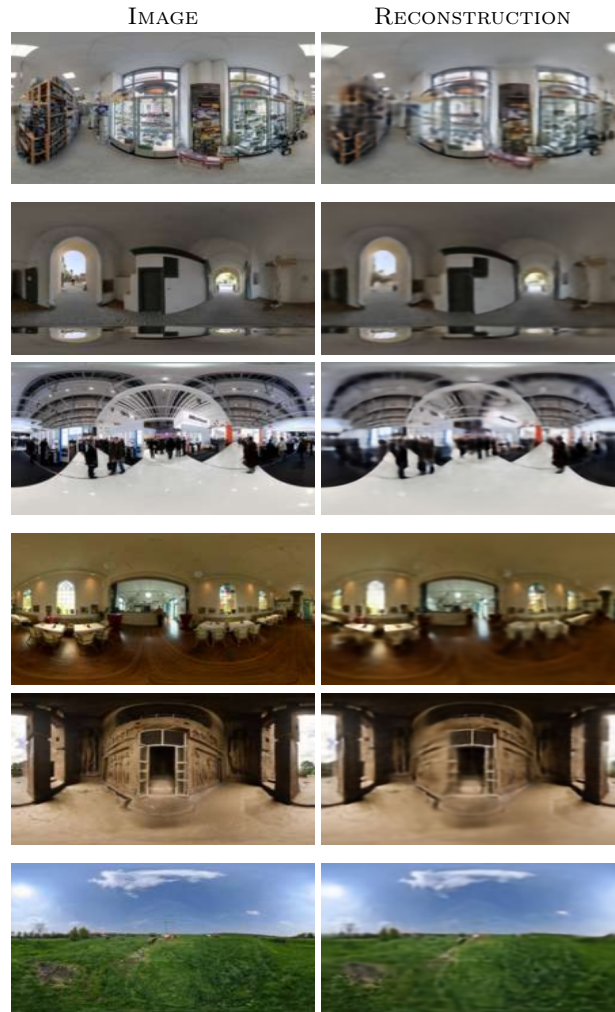


Fig. 8: Image reconstruction on SUN360: Figure shows the qualitative results of the image reconstruction task. Image size is 224×224 and 12×32^2 adaptive glimpses are used. For visualization, images were resized to the original aspect ratio.

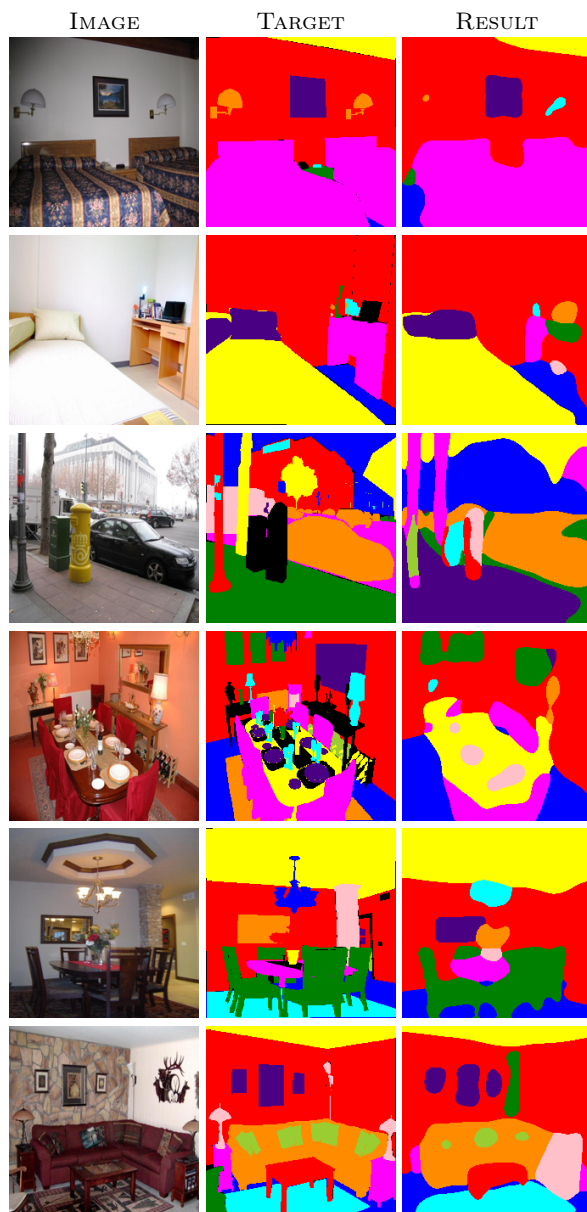


Fig. 9: Image segmentation on ADE20K: Figure shows the qualitative results of the image segmentation task. Image size is 224×224 and 12×48^2 adaptive glimpses are used.