

ZebraPose: Zebra Detection and Pose Estimation using *only* Synthetic Data

Supplementary Material

Overview

This is the supplementary material for ZebraPose. In Table 5, we provide a summary of the training and validation datasets. In there, we include the number of images for the training and validation sets, along with the types of animals contained in each dataset used in this work. Additionally, we present several qualitative results to further illustrate the performance of our models across different datasets and training configurations in Figure 7 for YOLO, and Figures 8 to 13 for ViTPose. All the data (synthetic and real), network weights, results, and code will be open-source.

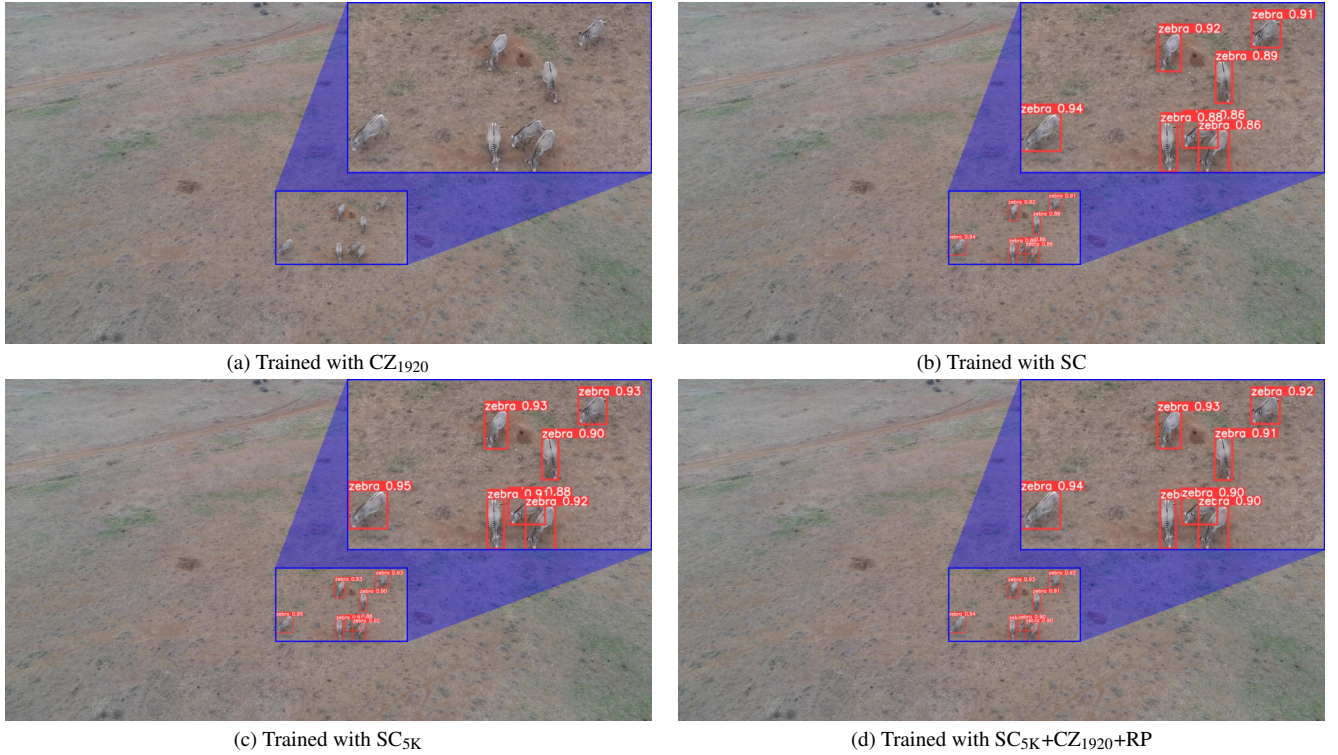


Figure 7. YOLOv5s results on images taken from [31] with images scaled to 1920px.

	Synthetic			Real, Common										Real, Aerial	
	SC [3]	SC _{5K} New	SpacNet [14]	A10	A10 _{OZ} [37]	A10 ₉₉	A36 [36]	A36 _{OZ}	Zebra-300 [16]	Zebra-Zoo [16]	TDH [8]	TDH ₉₉	CZ _[1920,640] [22]	R123	RP
Train	14401	23184	2640	7023	140	80	28457	960	—	—	8380	80	1916	—	720
Valid	3599	5798	360	995	20	19	7026	240	300	100	1772	19	85	104K	185
Test	—	—	—	1997	40	—	—	—	—	—	—	—	—	—	—
Animal	Zebras	Zebras	Zebras	Various	Zebras	Zebras	Various	Zebras	Zebras	Zebras	Horses	Horses	Zebras	Zebras	Zebras

Table 5. Datasets used in this work. We indicate the number of images in each train/validation set and the animal(s) included.

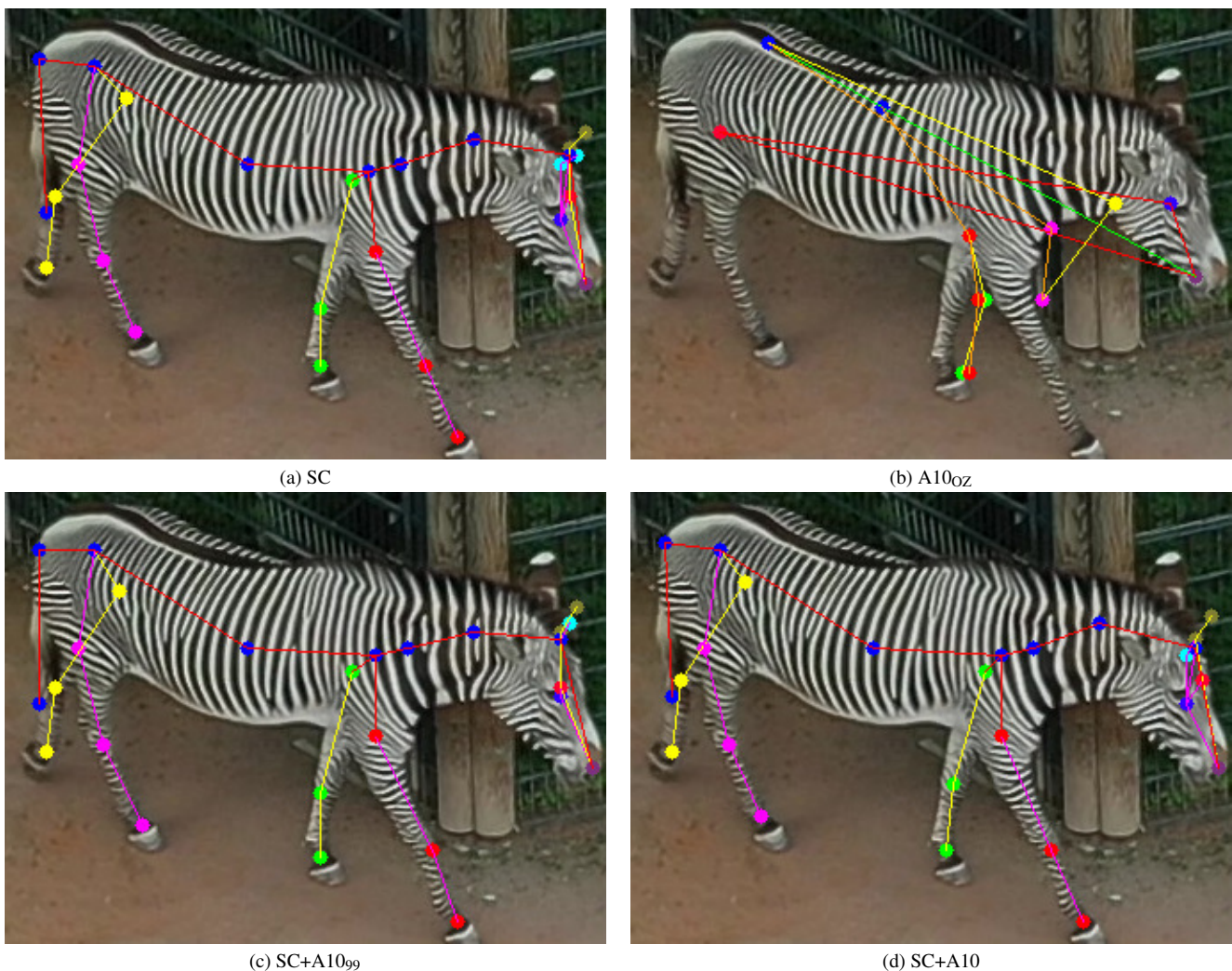
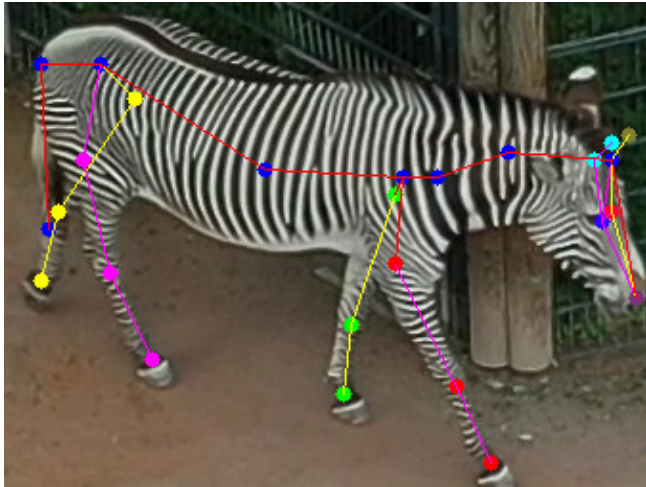
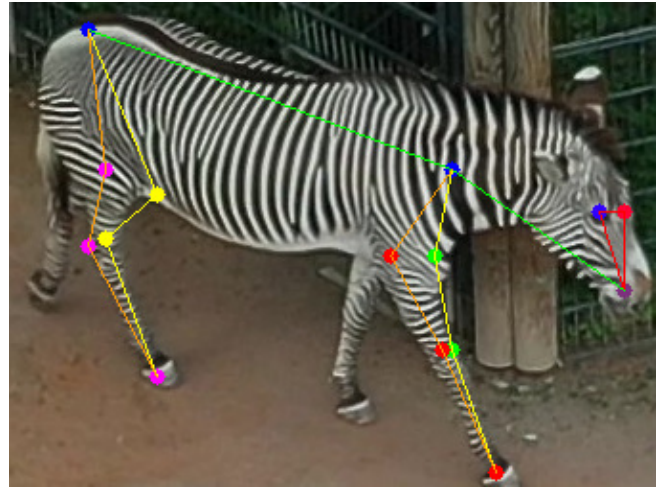


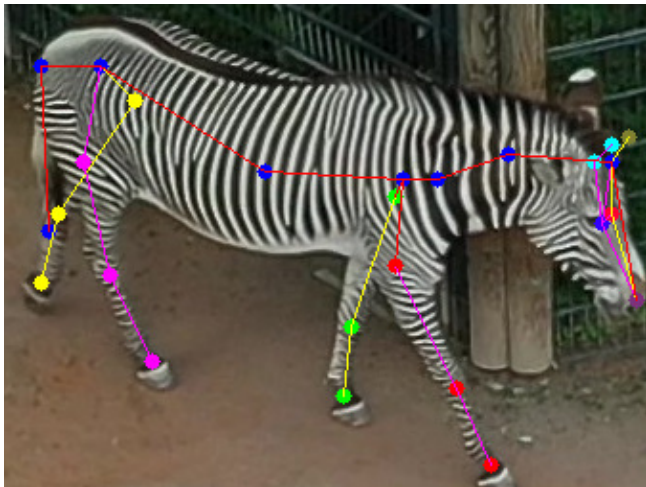
Figure 8. ViTPose+ trained on the specified dataset using a randomly initialized backbone and run on one of the images from the R123 dataset, manually cropped around the zebra *after* inference.



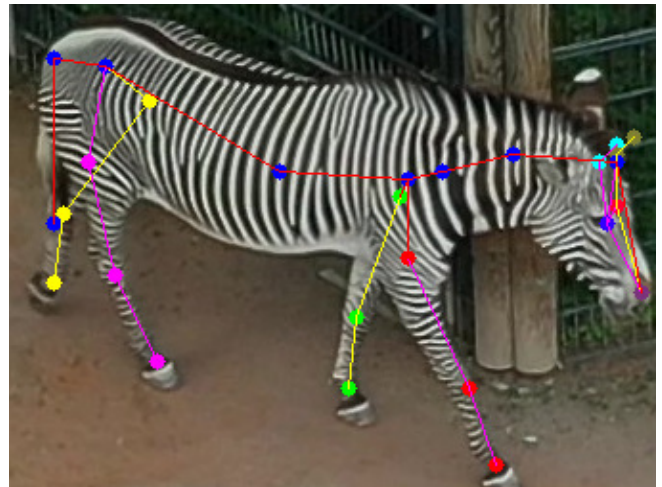
(a) SC



(b) A10oz



(c) SC+A10₉₉



(d) SC+A10

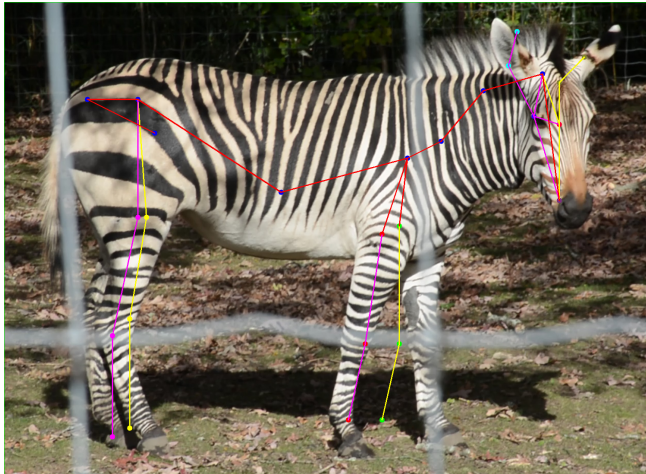
Figure 9. ViTPose+ trained on the specified dataset using a MAE pre-trained backbone and run on one of the images from the R123 dataset, manually cropped around the zebra *after* inference.



(a) SC



(b) A10oz



(c) SC+A10₉₉

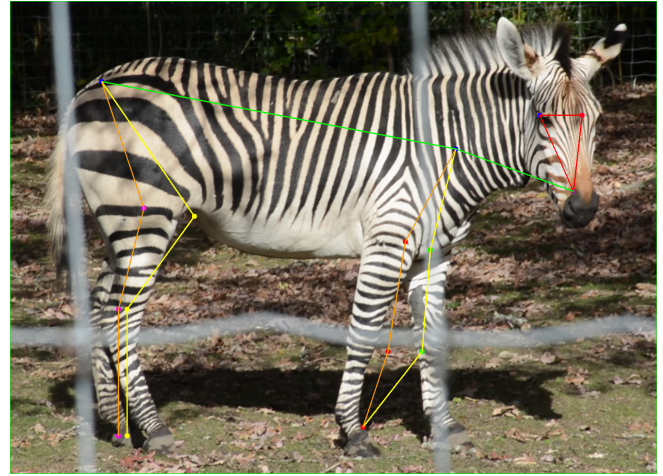


(d) SC+A10

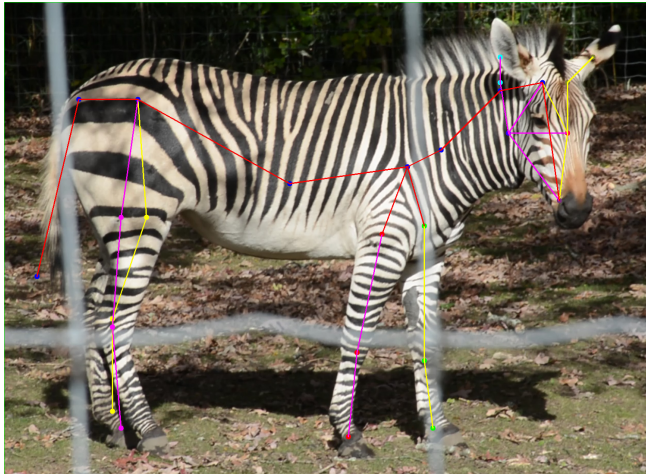
Figure 10. ViTPose+ trained on the specified dataset using a randomly-initialized backbone and run on one of the images from the Zebra-zoo dataset, manually cropped around the zebra *after* inference.



(a) SC



(b) A10oz

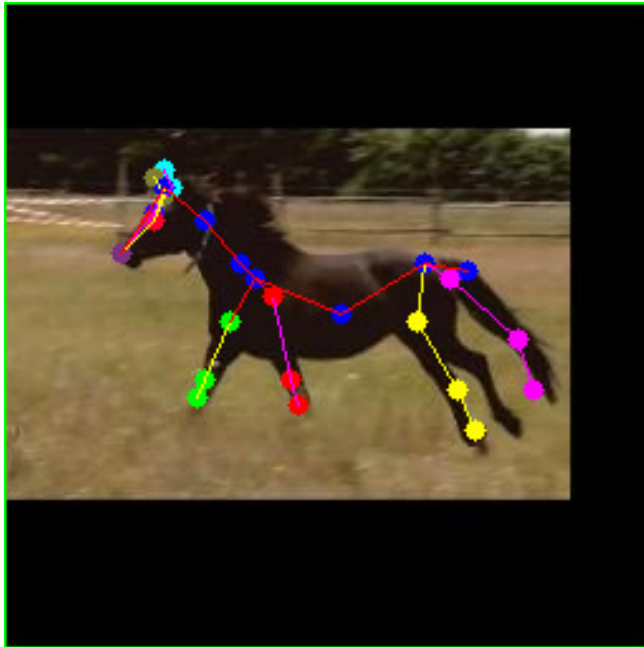


(c) SC+A10₉₉



(d) SC+A10

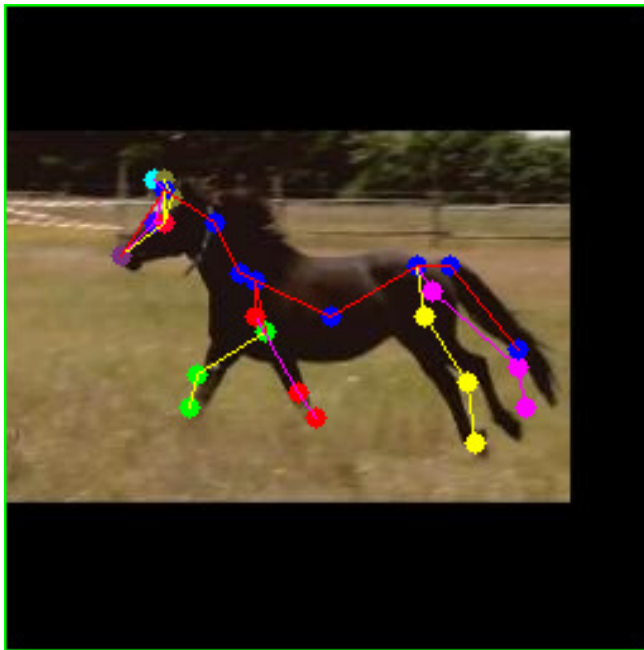
Figure 11. ViTPose+ trained on the specified dataset using a MAE pre-trained backbone and run on one of the images from the Zebra-zoo dataset, manually cropped around the zebra *after* inference.



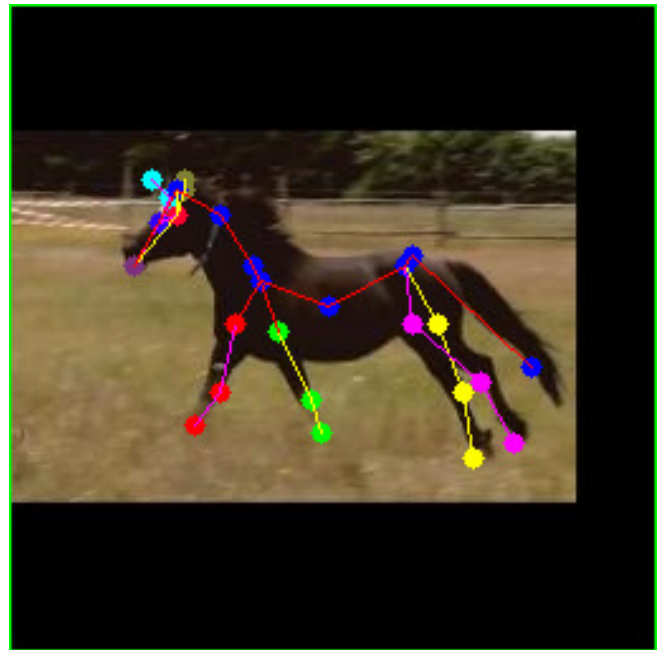
(a) SC



(b) A10oz

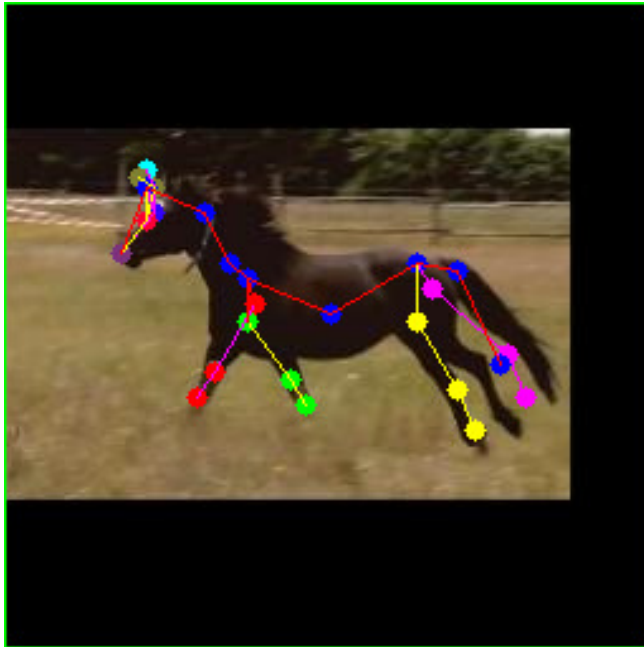


(c) SC+TDH₉₉

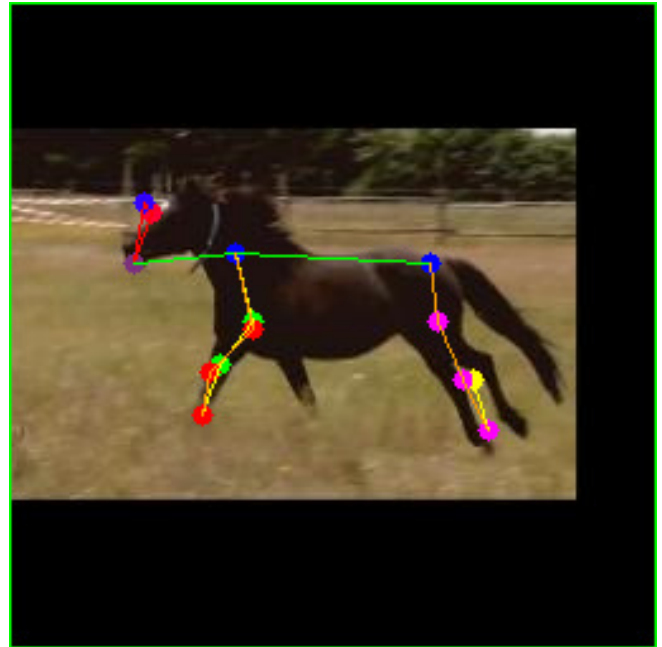


(d) SC+TDH

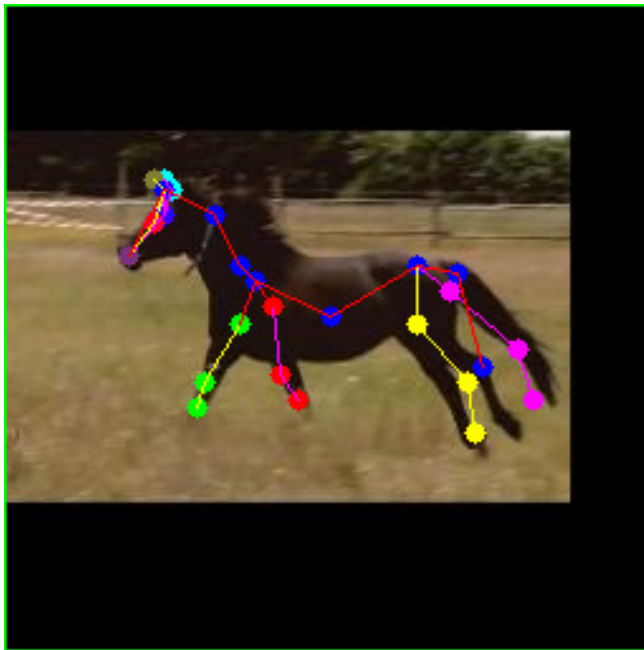
Figure 12. ViTPose+ trained on the specified dataset using a randomly-initialized backbone and run on one of the images from the TDH dataset shown as *processed* per dataset specifics.



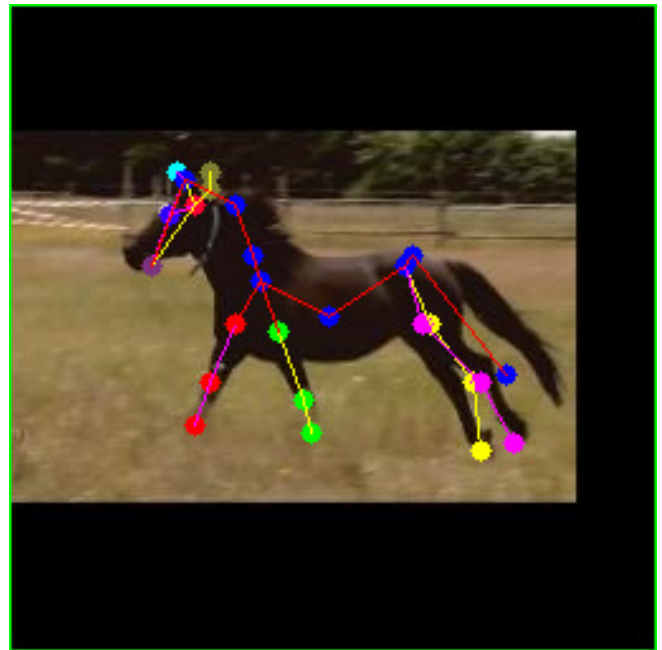
(a) SC



(b) A10oz



(c) SC+TDH₉₉



(d) SC+TDH

Figure 13. ViTPose+ trained on the specified dataset using a MAE pre-trained backbone and run on one of the images from the TDH dataset shown as *processed* per dataset specifics.