AnchorNet: A Weakly Supervised Network to Learn Geometry-sensitive Features For Semantic Matching

Supplementary material

David Novotny[1,2] Diane Larlus[2] Andrea Vedaldi[1]

[1]Visual Geometry Group
Dept. of Engineering Science, University of Oxford
{david,vedaldi}@robots.ox.ac.uk

[2]Computer Vision Group
Xerox Research Centre Europe
diane.larlus@xrce.xerox.com

1. Learning details

In this section we provide additional details about the learning protocol of AnchorNet. Training converges after visiting 4×10^4 training samples (for each class) in stage 1 and 1.2×10^4 samples in stage 2 (two days on a single GPU NVIDIA Tesla M40). The learning rate was fixed to a value of 10^{-2} with the minibatch size of 16 and the momentum set to the standard value of 0.0005. The training data were augmented as in [2].

The losses were balanced as follows. The weights of $\mathcal{L}_{\text{Discr}}$ and $\mathcal{L}^{\text{Aux}}_{\text{Discr}}$ were set to 1 and 10 respectively. The weight of $\mathcal{L}_R$ was set to a higher value of 10^6 which is necessary due to the inhibition of the gradient by the ℓ_2 normalization which takes place just before computing $\mathcal{L}_R$. The weights of $\mathcal{L}^{A,B}_{\text{Div}}$ and $\mathcal{L}_{\text{R}}$ were set to be as high as possible (10^5) such that $\mathcal{L}^{A,B}_{\text{Div}} \approx \mathcal{L}_{\text{R}} \approx 0$ are treated approximately as hard constraints. Importantly, $\mathcal{L}_R$ is optimized only during visiting positive samples as reconstructing the activations of negative samples would waste the capacity of the autoencoder. During the first training stage, we sample positive and negative samples with equal probability. Furthermore, during stage 2, we ensure that the distribution of positive samples is uniform over the set of 20 Pascal categories. This causes the positive samples from any given object category to be $20\times$ less frequent than the negative samples. Hence, in order to rebalance losses in stage 2, we decrease the weights of negative samples by a factor of 20. Due to the fact that the gradients from $\mathcal{L}^{A,B}_{\text{Div}}$ exhibit high magnitudes, we decrease the learning rate on the layers bellow the first autoencoder layer by a factor of 10^4 during the second stage.

2. Additional experimental results

Tables A and B provide an extension of Tables 1 and 2 from the paper. On top of the features already provided in Tables 1 and 2, we include more baseline features: res4c and res5c, which are extracted from the ResNet50 architecture and the features from Simon et al. [4]. [4] selects part-like convolutional feature channels using a mixture of constellation models; however, if two different aspects are detected in two images, the set of common features is too sparse for matching. Thus, we converted their output to dense descriptors for use in DSP and PF by 1) modifying the HC from the ResNet50 architecture by retaining their part-like channels across all aspects (denoted as **Constellation-HC**) and 2) by backpropagating the part-like channel activations to the input image as they do, and using the image-level activations as dense descriptors (**Constellation-BP**).

Additionally, to quantify the impact of the diversity losses $\mathcal{L}_{\text{Div}}$, we also report the performance of the features produced by the ANet-class method optimized without the diversity losses with DSP used as the matching algorithm (**DSP + ANet-class w/o $\mathcal{L}_{\text{Div}}$**).

We observe that the res4c, res5c features as well as all the variants of the constellation features perform on par with the hypercolumn features (HC). The apparent drop in performance of DSP + ANet-class w/o $\mathcal{L}_{\text{Div}}$ compared to DSP + ANet-class highlights the contribution of the diversity losses.

3. Additional qualitative results

Anchor filters. Figure 2 from the paper illustrates the anchor filters discovered by the ANet for the bird and dog classes. In Figures A and B, we illustrate anchor filters for the remaining classes, *i.e.* bicycle, bottle, bus, car, cat, chair, cow, dinning table, horse, motorbike, person, potted plant, sheep, sofa, and tv / monitor.

Segmentation transfer on PascalParts. Figure C complements Figure 4 from the paper and contains additional qualitative results for the segmentation transfer task on the

PASCAL Parts dataset.

Semantic matching on AnimalParts. Figure D complements Figure 6 from the paper and contains additional qualitative results for the semantic matching task on the AnimalParts dataset.

References

[1] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow. In *Proc. CVPR*, 2016. 3, 6, 7

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proc. CVPR*, 2016. 1

[3] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *Proc. CVPR*, 2013. 3, 6

[4] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proc. ICCV*, 2015. 1, 3

[5] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proc. CVPR*, 2015. 3

	mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	sofa	table	train	tv
Pairwise alignment methods																					
DSP + ANet-class	**0.45**	**0.31**	**0.49**	**0.32**	0.53	**0.75**	**0.51**	**0.47**	0.23	0.53	**0.37**	0.20	0.33	**0.41**	**0.22**	**0.46**	**0.45**	**0.77**	0.45	0.48	**0.74**
DSP + ANet-class w/o $\mathcal{L}_{\text{Div}}$	0.41	0.27	0.42	0.25	0.51	0.72	0.46	0.42	0.21	0.52	0.32	0.19	0.30	0.33	0.18	0.44	0.34	0.75	0.42	0.48	0.64
DSP + ANet	**0.45**	0.29	0.47	0.29	0.52	0.73	0.50	0.46	**0.25**	0.53	**0.37**	**0.21**	**0.34**	0.39	0.20	0.44	0.45	**0.77**	0.45	0.51	**0.74**
DSP + res4c	0.41	0.28	0.43	0.23	0.50	0.73	0.47	0.43	0.20	0.52	0.31	0.15	0.27	0.34	0.19	0.39	0.36	0.74	0.44	0.48	0.65
DSP + res5c	0.40	0.27	0.42	0.23	0.50	0.73	0.47	0.42	0.20	0.51	0.31	0.15	0.26	0.33	0.19	0.39	0.35	0.74	0.44	0.48	0.65
DSP + HC	0.41	0.29	0.45	0.24	0.51	0.73	0.48	0.44	0.20	0.52	0.32	0.16	0.28	0.35	0.19	0.39	0.37	0.74	0.44	0.48	0.67
DSP + SIFT [3]	0.39	0.25	0.46	0.21	0.48	0.63	0.50	0.45	0.19	0.48	0.30	0.14	0.26	0.35	0.13	0.40	0.37	0.66	0.37	0.48	0.62
DSP + Constellation-HC	0.40	0.28	0.42	0.23	0.50	0.73	0.47	0.42	0.20	0.52	0.31	0.15	0.27	0.34	0.19	0.38	0.36	0.74	0.44	0.48	0.65
DSP + Constellation-BP	0.40	0.27	0.41	0.23	0.50	0.73	0.46	0.42	0.20	0.51	0.31	0.15	0.26	0.33	0.18	0.38	0.35	0.73	0.44	0.47	0.64
Proposal Flow + ANet-class	0.43	0.26	0.43	0.28	**0.54**	0.71	0.50	0.45	0.24	**0.54**	0.32	**0.21**	0.28	0.35	0.21	0.45	0.40	0.74	**0.46**	0.50	0.70
Proposal Flow + ANet	0.42	0.26	0.41	0.26	0.53	0.70	0.49	0.45	**0.25**	**0.54**	0.31	0.19	0.28	0.31	0.17	0.43	0.39	0.74	0.44	**0.52**	0.69
Proposal Flow + res4c	0.42	0.27	0.44	0.26	0.54	0.70	0.50	0.45	0.23	0.53	0.32	0.18	0.28	0.33	0.17	0.44	0.39	0.74	0.45	**0.52**	0.66
Proposal Flow + res5c	0.39	0.23	0.34	0.25	0.53	0.70	0.47	0.43	0.22	0.52	0.30	0.18	0.26	0.27	0.17	0.41	0.38	0.73	0.45	0.49	0.60
Proposal Flow + HC	0.42	0.26	0.42	0.26	**0.54**	0.70	0.50	0.45	0.23	0.53	0.32	0.18	0.27	0.32	0.18	0.43	0.38	0.74	0.45	0.51	0.64
Proposal Flow + HoG [1]	0.41	0.25	0.45	0.23	**0.54**	0.70	0.49	0.44	0.19	0.53	0.30	0.16	0.25	0.35	0.16	0.41	0.35	0.74	0.44	0.50	0.63
Proposal Flow + Constellation-HC	0.40	0.26	0.39	0.25	0.53	0.68	0.48	0.43	0.21	0.52	0.30	0.17	0.26	0.31	0.15	0.42	0.37	0.72	0.44	0.50	0.62
Proposal Flow + Constellation-BP	0.39	0.25	0.38	0.23	0.53	0.68	0.47	0.41	0.20	0.51	0.29	0.16	0.25	0.30	0.15	0.41	0.35	0.71	0.43	0.49	0.60
Baseline: NoFlow	0.39	0.27	0.40	0.22	0.50	0.73	0.46	0.42	0.20	0.51	0.30	0.15	0.25	0.32	0.18	0.38	0.34	0.74	0.44	0.47	0.64
Collective alignment methods																					
FlowWeb [5]	0.43	0.33	0.53	0.24	0.51	0.72	0.54	0.51	0.20	0.52	0.32	0.15	0.29	0.45	0.19	0.41	0.39	0.73	0.41	0.51	0.68

Table A: Weighted IoU for pairwise **semantic part matching** (not to be confused with object or part detection or segmentation) on PASCAL Parts. The methods that use our proposed features are in **bold**.

	mean	aero	bike	boat	bottle	bus	car	chair	mbike	sofa	table	train	tv
Pairwise alignment methods													
DSP + ANet-class	**0.24**	**0.23**	0.28	**0.06**	0.38	**0.44**	**0.39**	**0.14**	**0.19**	0.16	0.11	0.13	**0.41**
DSP + ANet-class w/o $\mathcal{L}_{\text{Div}}$	0.17	0.19	0.18	0.06	0.31	0.31	0.18	0.10	0.13	0.12	0.08	0.12	0.24
DSP + ANet	0.23	0.22	0.25	0.06	0.35	0.42	0.34	**0.14**	0.17	**0.17**	**0.13**	**0.14**	0.40
DSP + HC	0.20	0.20	0.23	0.05	**0.39**	0.36	0.25	0.10	0.15	0.12	0.10	0.12	0.28
DSP + res4c	0.19	0.20	0.22	0.05	0.39	0.35	0.24	0.10	0.14	0.11	0.09	0.12	0.27
DSP + res5c	0.17	0.19	0.19	0.05	0.38	0.32	0.19	0.09	0.13	0.11	0.08	0.11	0.25
DSP + SIFT [3]	0.18	0.17	**0.30**	0.05	0.19	0.33	0.34	0.09	0.17	0.12	0.09	0.12	0.18
DSP + Constellation-HC [4]	0.18	0.20	0.21	0.05	0.39	0.33	0.20	0.10	0.13	0.12	0.09	0.12	0.26
DSP + Constellation-BP [4]	0.17	0.19	0.19	0.05	0.39	0.32	0.19	0.10	0.12	0.12	0.08	0.12	0.25
Proposal Flow + ANet-class	0.17	0.17	0.21	0.05	0.25	0.26	0.27	0.10	0.14	0.12	0.07	0.10	0.24
Proposal Flow + ANet	0.16	0.16	0.19	0.05	0.22	0.26	0.25	0.10	0.12	0.11	0.05	0.12	0.23
Proposal Flow + HC	0.16	0.17	0.21	0.05	0.23	0.27	0.24	0.09	0.13	0.12	0.05	0.11	0.20
Proposal Flow + res4c	0.17	0.19	0.24	0.05	0.23	0.28	0.27	0.09	0.15	0.12	0.05	0.13	0.21
Proposal Flow + res5c	0.11	0.13	0.11	0.04	0.21	0.21	0.19	0.07	0.08	0.08	0.05	0.09	0.14
Proposal Flow + HoG [1]	0.17	0.20	0.26	0.05	0.20	0.31	0.29	0.10	0.17	0.13	0.05	0.13	0.21
Proposal Flow + Constellation-HC [4]	0.14	0.18	0.17	0.04	0.19	0.25	0.20	0.08	0.12	0.10	0.05	0.10	0.17
Proposal Flow + Constellation-BP [4]	0.13	0.16	0.15	0.04	0.19	0.25	0.18	0.07	0.10	0.10	0.06	0.10	0.17
Baseline: NoFlow	0.17	0.18	0.17	0.05	**0.39**	0.31	0.17	0.09	0.12	0.11	0.07	0.11	0.24
Collective alignment methods													
FlowWeb [5]	0.26	0.29	0.41	0.05	0.34	0.54	0.50	0.14	0.21	0.16	0.04	0.15	0.33

Table B: PCK ($\alpha = 0.05$) for semantic keypoint transfer on the 12 rigid classes of the PASCAL Parts dataset.

Figure A: Example anchor filters discovered by the AnchorNet for the bicycle, bottle, bus, car, cat, chair, cow, dining table, horse classes.

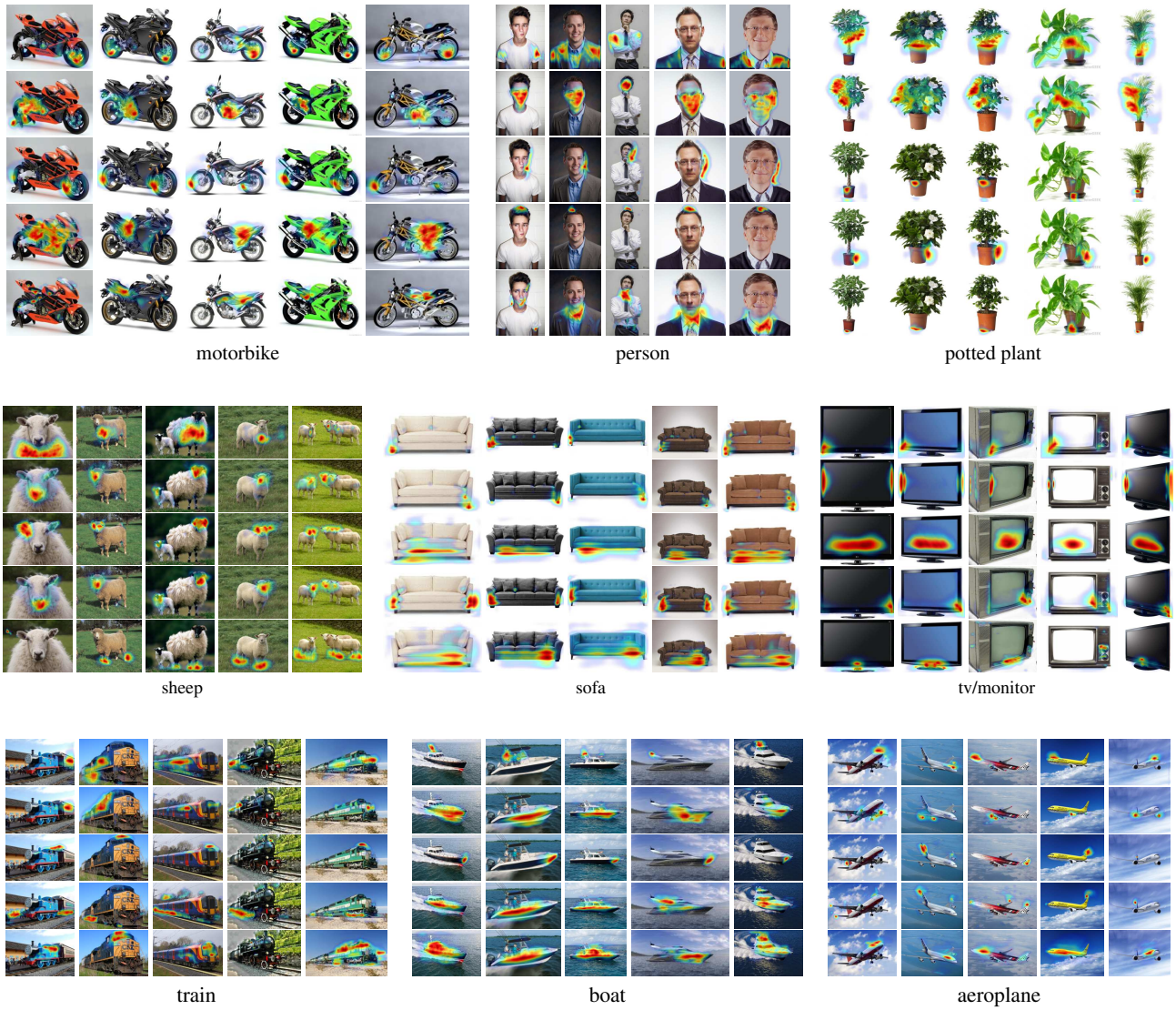

Figure B: Example anchor filters discovered by the AnchorNet for the motorbike, person, potted plant, sheep, sofa, tv/monitor, train, boat, aeroplane classes.

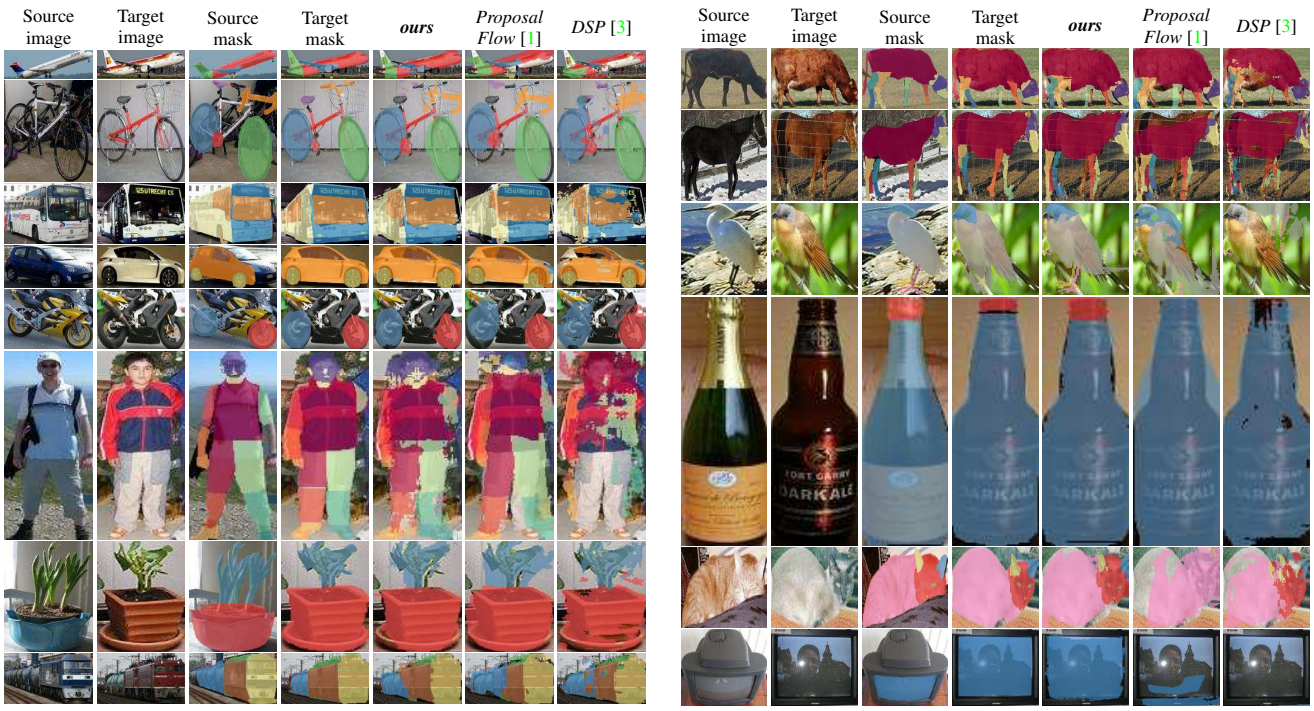

Figure C: **Segmentation mask transfer** on PASCAL Parts for DSP+ANet (ours), Proposal Flow + HoG, and DSP + SIFT.

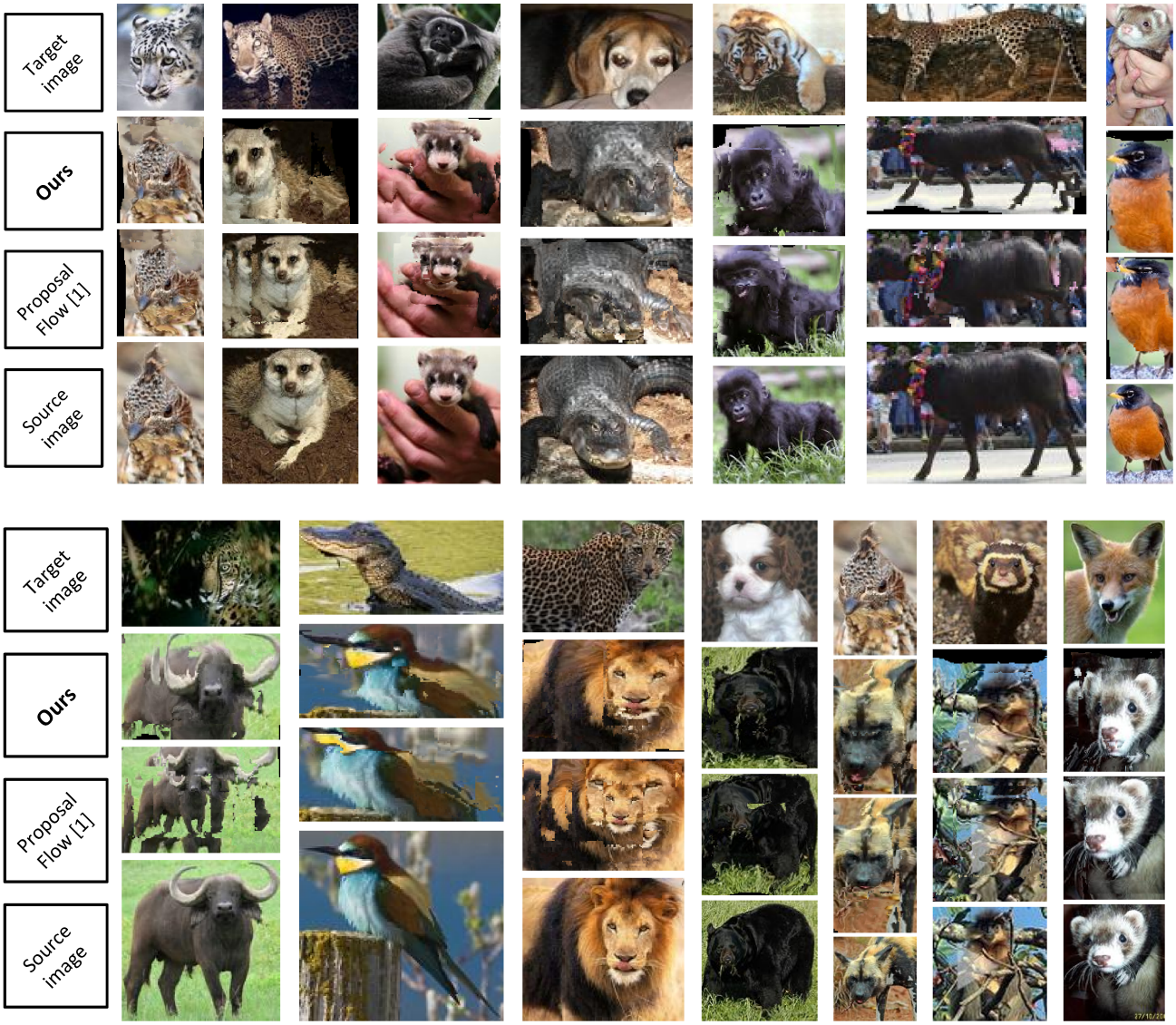

Figure D: **Cross-class alignments** on the AnimalParts dataset. Given a target (top row) and source images (bottom row) we establish semantic correspondences between parts of animal classes. The alignment warps the source image into the target image. We compare Proposal Flow + ANet (ours - 2nd row) and Proposal Flow + HoG [1] (3rd row).