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

UNIVERSITY OF OXFORD — NAVER LABS EUROPE

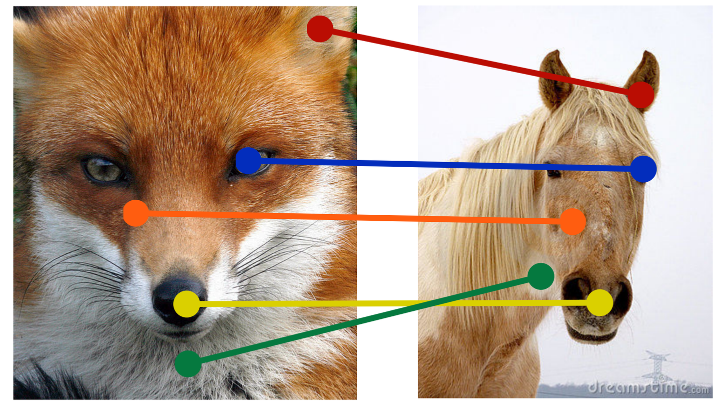
David Novotny[1,2], Diane Larlus[2], Andrea Vedaldi[1]

[1] Visual Geometry Group, University of Oxford, UK
[2] Naver Labs Europe, Grenoble, France

Introduction

The task:

Semantic matching



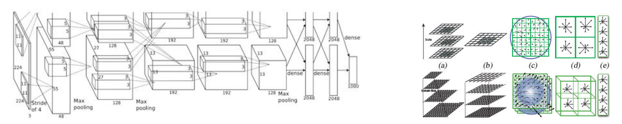
Given a pair of semantically related objects
=> estimate matches between corresponding parts

Motivation:

Fully supervised approaches [4,6] require expensive annotations / synthetic datasets => we target weak supervision

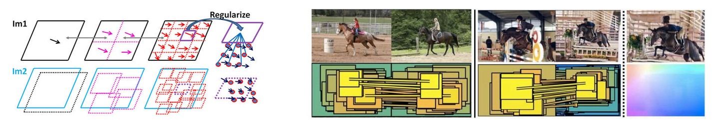
Weakly supervised approaches:

Step 1. Extract pixel-wise descriptors
> Pretrained deep features, HoG, ...



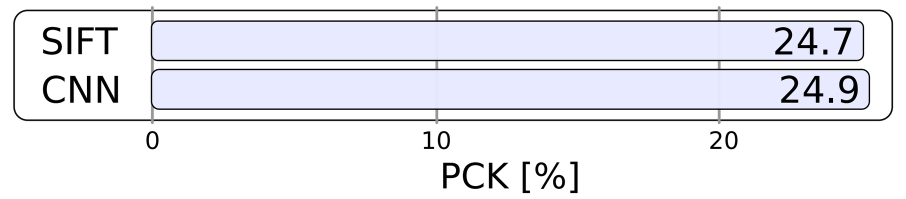
The main focus of the paper

Step 2. Use a matching algorithm
> DSP [1], Proposal Flow [2], SIFT Flow [3], ...



Pixel-wise descriptors - deep vs. engineered features:

Semantic matching accuracy on Pascal VOC [5]



=> deep features on par with engineered ones

Deep features trained with a **global classification loss**
=> attention to most discriminative regions
=> invariance to geometry of the input

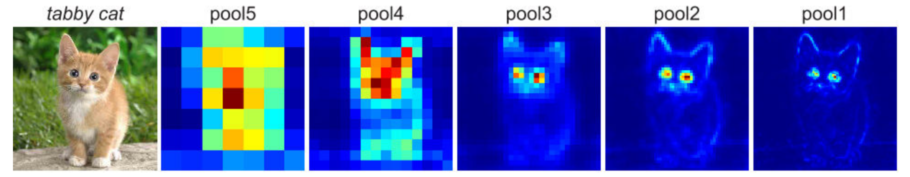


image courtesy of [10]

=> **Main challenge:** Design the training s.t. features are not invariant to geometry

Proposed approach - overview:

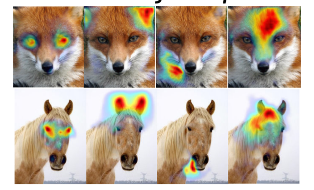
Train

Given a large dataset with object category **image level labels**

e.g. horse & fox images

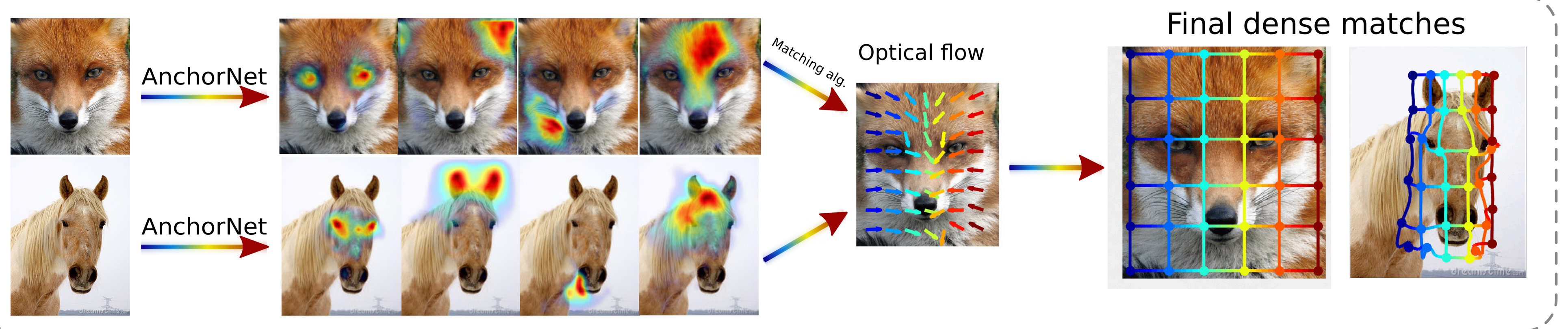
AnchorNet training

Learn distinct features of the object categories AKA *"object parts"*



Use the features within a matching algorithm

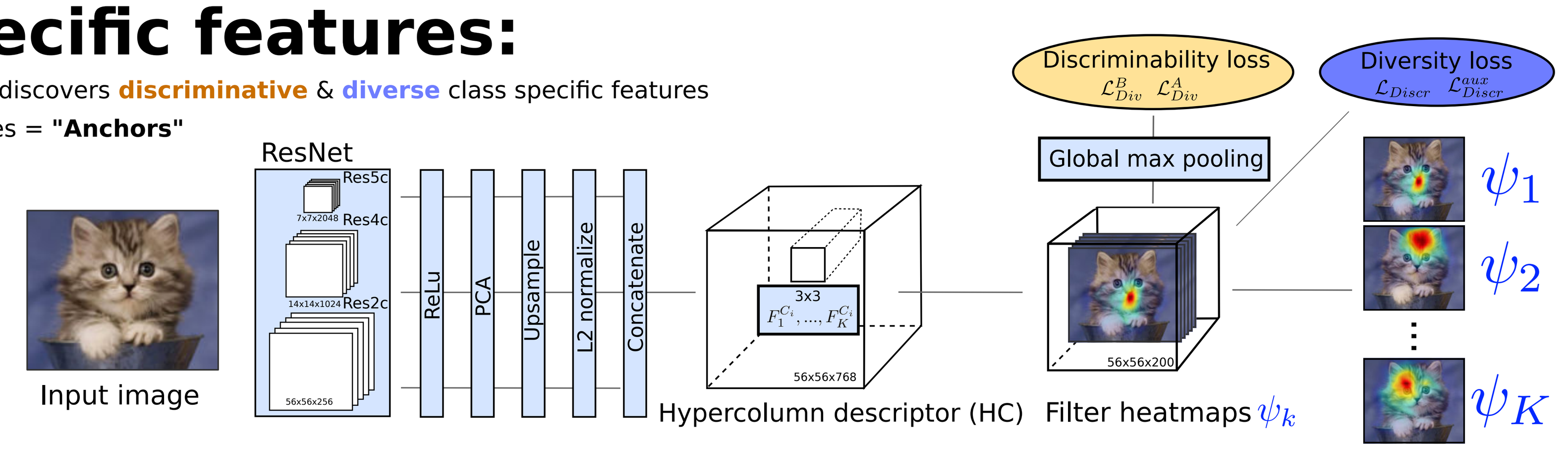
Test



AnchorNet architecture

Class specific features:

AnchorNet automatically discovers **discriminative** & **diverse** class specific features
=> discovered features = **"Anchors"**



Discriminability loss

$$\mathcal{L}_{\text{Discr}} = -y_I \sum_{k=1}^{K} \max \psi_k$$

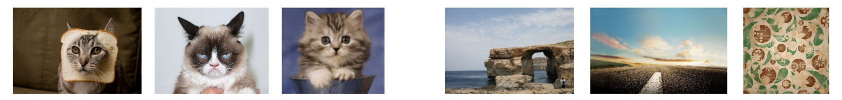
> identifies the representative of each filter by global max pooling
> increases response in positive images & decreases for negatives

Diversity loss

$$\mathcal{L}_{\text{Div}} = \sum_{i \neq j} \left\| \frac{\langle \psi_i, \psi_j \rangle}{\|\psi_i\|_F \|\psi_j\|_F} \right\|^2$$

> enforces orthogonal responses of the filters

Positive images — Negative images

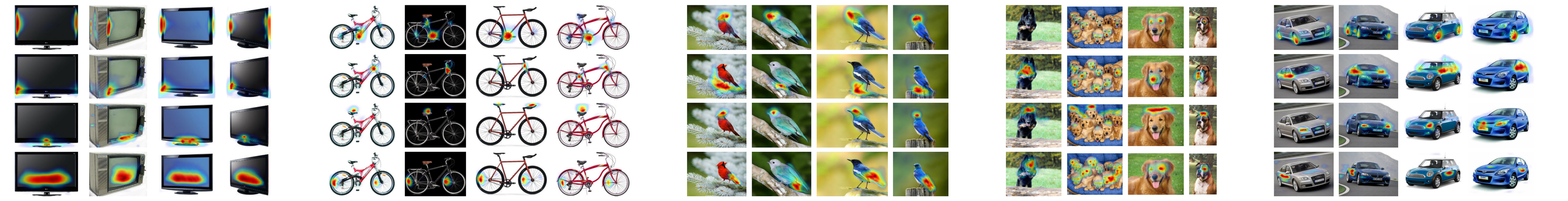


$y_I = 1$ — $y_I = -1$

Training **only** with Discriminability loss — Training with Discriminability & Diversity losses

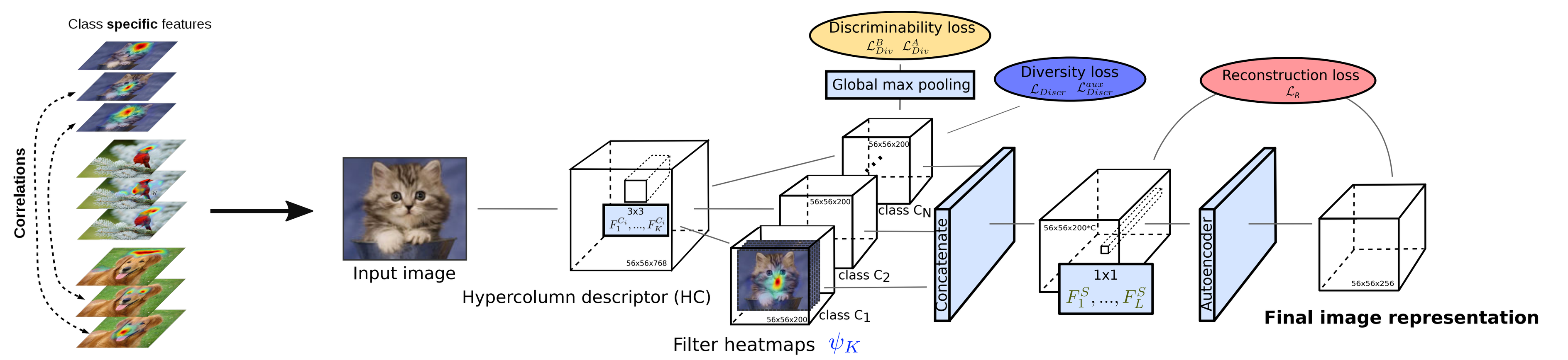
NO diversity loss => redundant features — **WITH diversity loss => full object coverage**

Example learned class specific filters



Class agnostic features:

Class specific filters exhibit correlations => **use an autoencoder to compress the features shared between classes**



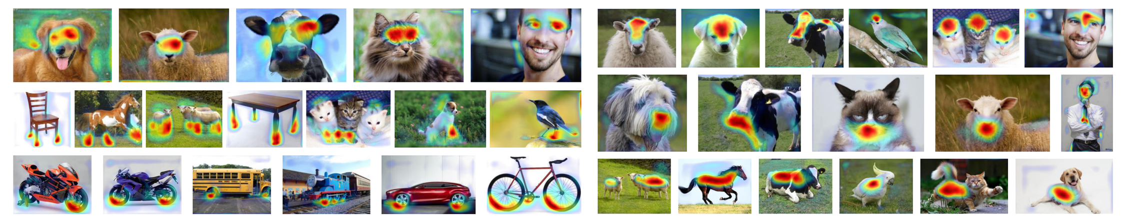
Reconstruction loss

$$\mathcal{L}_R = \mathcal{D}\left(\Gamma,\ (F^S)^T F^S \Gamma\right)$$

$\Gamma = \text{stack}(\psi_1^1, \ldots, \psi_K^C)$ - concatenation of the class-spec. features ψ_k

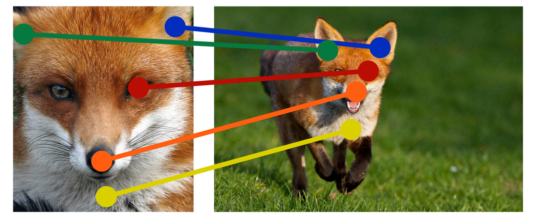
> compresses the correlated class specific features
> produces a class agnostic representation

Example learned class agnostic filters



Experiments

Semantic matching:



Given a pair of images of **the same object category**
=> estimate matches between corresponding parts

Evaluated approach

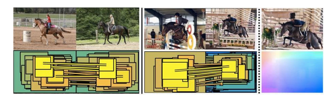
Step 1. Extract pixelwise descriptors
> AnchorNet features
  > ANet-class ... class specific features
  > ANet ... class agnostic features
> SIFT, HoG, Hypercolumns

Step 2. Match descriptors using a matching alg.
> DSP [1]
> Proposal Flow [2]

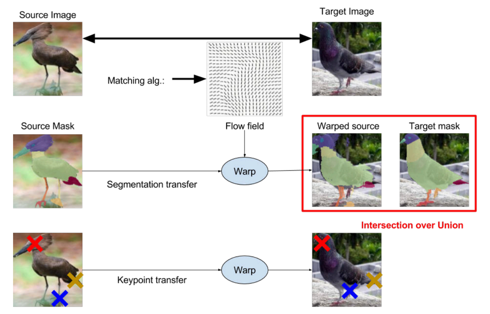
Benchmarks

Pascal Parts [7]
> part segmentation masks + keypoint annotations
> the 20 Pascal VOC classes

PF Dataset [2]
> dense correspondence annotations
> 6 different object classes



Evaluation procedure



Pascal Parts - segmentation transfer - *IoU*

IoU : Intersection-over-Union AKA Jaccard index

	mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	sofa	table	train	tv
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	**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 + SIFT [1]	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
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 + HoG [2]	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

Pascal Parts - keypoint transfer - *PCK@0.05* [%]

PCK ... percentage of correctly transfered keypoints

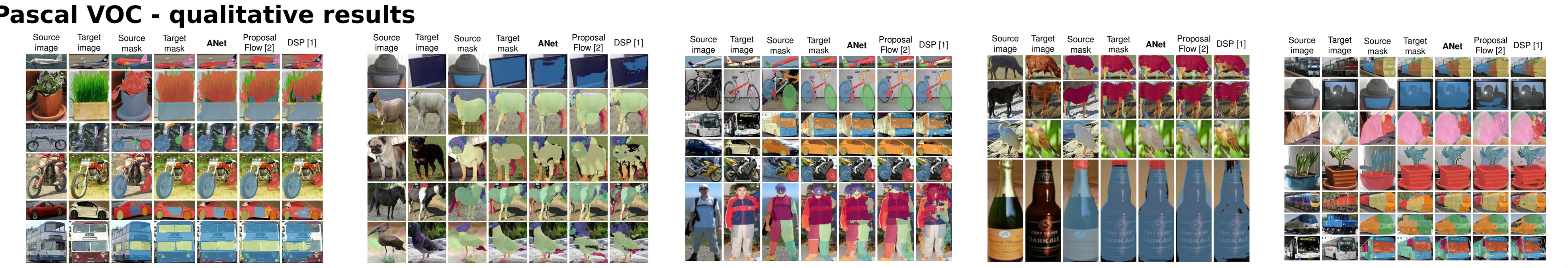
	mean	aero	bike	boat	bottle	bus	car	chair	mbike	sofa	table	train	tv
DSP + ANet-class	**0.24**	**0.23**	0.28	**0.06**	**0.38**	**0.44**	**0.39**	**0.14**	**0.19**	0.16	0.11	0.13	**0.41**
DSP + ANet	0.23	0.22	0.25	0.06	0.35	0.42	0.34	**0.14**	0.17	**0.17**	**0.13**	**0.14**	0.40
DSP + SIFT [1]	0.18	0.17	**0.30**	0.05	0.19	0.33	0.34	0.09	0.17	0.12	0.09	0.12	0.18
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 + HoG [2]	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

PF Dataset

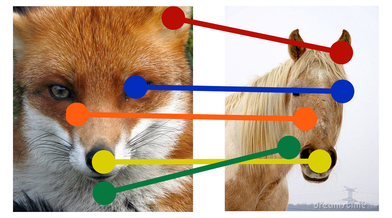
Matching \ Feature	AuCs for PCR		
	ANet-class	**ANet**	HoG [2]
NAM: baseline	**0.41**	0.36	0.29
LOM: Proposal Flow	**0.46**	0.43	0.43

→ State-of-the-art performance on the segmentation transfer task
→ AnchorNet features improve performance of matching algorithms
→ AnchorNet vs AnchorNet-class perform on par => successfull conversion from class specific to class agnostic features
→ PF Dataset - naive matching of AnchorNet features similar to matching engineered features with a sophisticated algorithm

Pascal VOC - qualitative results



Cross-class semantic matching:

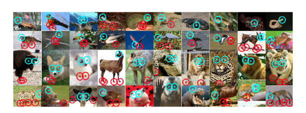


Given a pair of images of **related object categories**
=> estimate matches between corresponding parts

Benchmarks

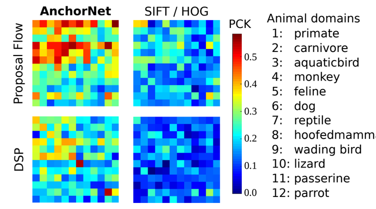
Pascal Parts [7]
> related object classes share part segmentations
> e.g. "car" and "bus" classes share "wheel", "door", "window", ... parts
> evaluation of segmentation transfer between images of meaningful classes
> low number of categories, large number of shared parts

Animal Parts [8]
> images from the ImageNet dataset annotated with "eye" and "foot" keypoints
> evaluation of keypoint transfer between images from different animal domains
> large number of categories, low number of parts



Animal Parts - Cross class keypoint matching accuracy - *PCK*

Transfer accuracy between animal domains



Animal domains
1: primate
2: carnivore
3: aquaticbird
4: monkey
5: feline
6: dog
7: reptile
8: hoofedmammal
9: wading bird
10: lizard
11: passerine
12: parrot

Mean PCK over all matches

Matching Alg.	DSP		Proposal Flow	
Feature	**ANet**	SIFT [1]	**ANet**	HoG [2]
PCK (α = 0.05)	0.11	0.06	**0.13**	0.06
PCK (α = 0.1)	0.24	0.12	**0.32**	0.18

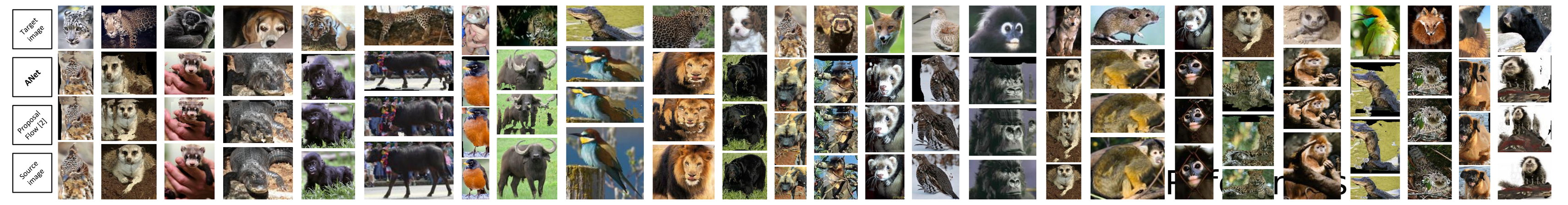
Pascal Parts - Cross-class segmentation transfer - *IoU*

> Mean IoU over transfered segmentations of shared parts

Source class ↓ Target class	mean	bicycle ↓ mbike	mbike ↓ bicycle	bus ↓ car	car ↓ bus	bus ↓ car	dog ↓ cat	cat ↓ dog	sheep ↓ dog	dog ↓ sheep	horse ↓ cow	cow ↓ horse	sheep ↓ cow	cow ↓ sheep
DSP + ANet	**0.37**	**0.35**	**0.45**	**0.52**	**0.35**	0.36	**0.25**	0.25	**0.34**	**0.27**	**0.31**	**0.47**	**0.37**	**0.58**
DSP + SIFT [1]	0.29	0.28	0.40	0.40	0.27	0.30	0.16	0.16	0.20	0.19	0.26	0.31	0.28	0.50
Proposal Flow + ANet	0.35	0.32	0.38	0.50	0.32	**0.37**	0.23	**0.27**	0.30	0.25	0.29	0.41	0.32	0.53
Proposal Flow + HOG [2]	0.31	0.30	0.43	0.48	0.30	0.35	0.19	0.21	0.22	0.19	0.25	0.37	0.29	0.50

→ AnchorNet features bring significant improvement over considered baseline features
→ State-of-the-art performance on both datasets

Animal Parts - qualitative results



Conclusions

→ Proposed **AnchorNet** - a weakly supervised architecture for learning geometry-sensitive features
→ The learned features are invariant to appearance making them suitable for semantic matching tasks
→ Experimentally verified that the features improve performance of existing matching algorithms
→ State-of-the-art performance on semantic matching and on novel cross-class semantic matching task

References

[1] Kim et al. "Deformable spatial pyramid matching for fast dense correspondences." CVPR 2013.
[2] Ham et al. "Proposal flow." CVPR 2016.
[3] Liu et.al. "Sift flow: Dense correspondence across scenes and its applications." CVPR 2011.
[4] Choy et al. "Universal correspondence network." NIPS 2016.
[5] Long et al. "Do convnets learn correspondence?" NIPS" 2014.
[6] Zhou et al. "Learning dense correspondence via 3d-guided cycle consistency." CVPR 2016.
[7] Chen et al. "Detect what you can: Detecting and representing objects using holistic models and body parts." CVPR 2014.
[8] Novotny et al. "I have seen enough: Transferring parts across categories." BMVC 2016.
[9] Hariharan et al. "Hypercolumns for object segmentation and fine-grained localization." CVPR 2015.
[10] Zhang et al. "Top-down neural attention by excitation backprop." ECCV 2016.