

UNIVERSITÀ
DEGLI STUDI
DI PADOVA



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



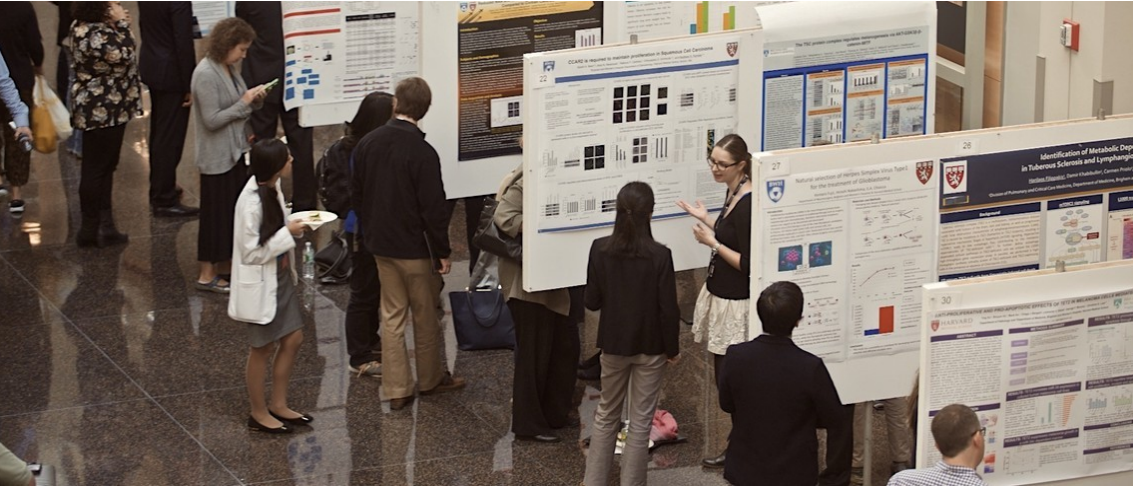
IAS-Lab

Intelligent Autonomous
Systems Laboratory

3D Data Processing Poster Session

Alberto Pretto

Poster Session



Poster Session

Groups: 2 students

When: June 11 2024, h.12:30 – 14:30

Where: in the corridor in front of the exit door of Ve classroom (third floor of DEI/G)

 **Day and location to be confirmed!**

Papers

We will shortly publish a selection of recent papers about:

- Stereo/multi-view stereo
- 3D reconstruction/Structure from Motion
- 3D SLAM/Visual SLAM
- Local/global 3D features
- Point cloud preprocessing/semantic segmentation
- 3D object pose estimation
- 3D Shape registration
- 3D sensors
- Neural fields
- ...

You will have a couple of days to checkout them, then each paper will be assigned to only one group with a "first come-first served" policy

Selection via moodle, we will communicate the time/date shortly

Poster Session

- The poster session will be **public**, so you will have the opportunity to present your poster to other interested students, PhD students, or researchers from DEI.
- A poster presentation is typically 6-7 minutes long (around 3 minutes each student), and it is more informal and interactive compared with the classical oral presentations.
- All the posters are set up at once, and each presenter is expected to stand with their poster for the entirety of the session to answer questions from attendees.

Poster Evaluation

My collaborators and I will review each poster, listening to your presentation and asking questions.

Poster Layout

Format: The poster should be prepared as a single face, landscape or portrait, A0 sheet, and saved in **PDF** format.

Tools: You may use Microsoft Office Power Point, LibreOffice Impress, Google Slides, ...

Poster Submission

- Through the e-learning website (a submission link will be shared soon) by **June 9 h. 24.** I will print the posters you submitted, so on the day of presentation I'll bring the posters
- You may also bring your poster on June 11, but in this case you will have to pay for the printing of the poster 😊...

What is a Research Poster?

- Posters should summarize the paper **concisely** and **attractively** to help advertise it and generate discussion.
- The poster should report **brief and clear text blocks** mixed with easy to understand **tables, graphs, pictures**.
- The poster should support the presentation and the interaction with the audience.

What Makes a Good Poster?

- Important information should be readable from 3 m
- Word count: 500 to 1000 words
- Text should be clear and direct → subject, verb and object complement!
- Use bullets, numbering, and boxes.
- Consistent, clean and attractive layout → **Effective use of graphics, color and fonts.**

"Bad" Poster



Multi-Image Semantic Matching by Mining Consistent Features

Qianqian Wang, Xiaowei Zhou, Kostas Danilidis
Zhejiang University, University of Pennsylvania



INTRODUCTION

- Problem: **Multi-image Semantic Matching**, i.e. finding feature correspondences across different object instances or scenes, in a large collection of images (in the order of thousands or more).
- Applications: Object-class model reconstruction, Automatic landmark annotation.
- Standing challenges in semantic and multi-image matching:
 - Repeatable feature point detection is an open problem.
 - Simultaneous optimization of cycle+geometric consistency.
 - Existing methods are computationally expensive.

PROPOSED SOLUTION

- Key idea: Identify and match only a **sparse set of highly repeatable features** in the image collection.
- In this way, the proposed method is able to explicitly prune nonrepeatable features and it is also highly scalable to handle thousands of images.
- Dense correspondences can be later achieved by interpolation.
- In addition, a **low-rank constraint** was imposed as an **efficient** way to ensure geometric consistency over the whole image collection.

PRELIMINARIES

- Given $1 < l < n$ images to match and p_i feature points per image, pairwise feature correspondences for a pair (i, j) can be represented by a partial permutation matrix $P_{ij} \in \{0, 1\}^{p_i \times p_j}$.
- Pairwise matching:** Individual pairs P_{ij} can be estimated using the Hungarian algorithm or approximated graph matching algorithms. Said estimates will be denoted by $P_{ij} \approx W_{ij} \in \mathbb{R}^{p_i \times p_j}$.
- Cycle consistency:** Constraint used in the multi-image case. For any triplet (i, j, k) , it must hold: $P_{ij} \cdot P_{jk} = P_{ik}$. Given $X_i \in \{0, 1\}^{p_i \times r}$, the correspondence between image i and the universe of unique features in the collection. The set $\{P_{ij} | P_{ij} \cdot X_i\}$ is cyclically consistent iff P can be factorized as XX^T , where $X \in \{0, 1\}^{n \times r}$ and:

$$P = \begin{bmatrix} P_{11} & \dots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \dots & P_{nn} \end{bmatrix}, X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

PROPOSED METHODS

- Cycle consistency with feature selection:** Instead of solving over the whole universe of features, consider the k most repeatable features (k is a small user-defined value).
- Estimate X_i through minimization subject to a **sparsity constraint** dependent on k : $\min_X \frac{1}{2} \|W - XX^T\|_F^2$, s.t. $X_i \in \{0, 1\}^{p_i \times k}$.
- Where, $W \in \mathbb{R}^{n \times n}$ is the concatenation of W_{ij} , i.e. the initial (non-consistent) estimation of P computed by pairwise solvers.
- Geometric constraint:** Spatial relationships (ex. nose below eyes). **Low rank constraint** on the measurement matrix M across scene frames (rank 4 under orthographic projection in rigid scenes [1]).
- Assume n different images of the same object class as n frames of a non-rigid scene. M still approximated by a low-rank (r) matrix.
- $M \in \mathbb{R}^{2n \times k}$ built with the coordinates of the k feature points selected from each image $M_i = C_i X_i$, where $C_i \in \mathbb{R}^{2 \times p_i}$.
- Geometric constraint imposed by minimizing: $\min_{X, Z} \frac{1}{2} \sum_{i=1}^n \|C_i X_i - Z_i\|_2^2$, s.t. $\text{rank}(Z) \leq r$, where $Z \in \mathbb{R}^{2 \times k}$ is an auxiliary variable.

FORMULATION

- Combining cycle and geometric consistency terms gives the final optimization problem:
- $$\min_{X, Z} \frac{1}{2} \|W - XX^T\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^n \|C_i X_i - Z_i\|_2^2$$
- s.t. $X_i \in \{0, 1\}^{p_i \times k}$ $\Lambda \text{rank}(Z) \leq r$
- Where λ controls the impact of the geometric constraint.

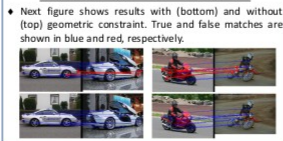
PRACTICAL IMPLEMENTATION

- Replace X in the first term with an auxiliary variable $Y \in \mathbb{R}^{n \times k}$, i.e. real matrix representing a permutation matrix.
 - Add a soft constraint to push Y towards X .
- $$\frac{1}{2} \|W - YY^T\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^n \|C_i X_i - Y_i\|_2^2 + \frac{\rho}{2} \|X - Y\|_F^2$$
- s.t. $X_i \in \{0, 1\}^{p_i \times k}$ $\Lambda \text{rank}(Z) \leq r$ $\forall Y \in C$
- Where, $C: 0 \leq Y \leq 1, 0 \leq Y_i \leq 1, Y_i^T = 1$
- Apply block coordinate descent, i.e., alternately updating one variable X, Y, Z while fixing the others.
 - Y updated via projected gradient descent.
 - Each X_i updated via the Hungarian algorithm.
 - Z updated via singular value decomposition.
 - Empirically set λ as 1, r as 4, ρ as sequence (1, 10, 100).
 - Reliably initialize Y by ignoring geometric constraint, $\min_Y \frac{1}{2} \|W - YY^T\|_F^2$ s.t. $Y \in C$. Discretize Y to init X .

BENCHMARK 1: MULTI-GRAPH MATCHING

- Goal:** Feature points are annotated for every image, but their correspondences need to be estimated.
- Datasets:** CMU (hotel, house), and WILLOW Object Class (car, duck, motorbike, face, winebottle).
- Implementations:** 3 tested: Ours- and Ours, without and with geometric constraint respectively, W obtained from the Hungarian algorithm. Whereas, Ours+ employed the graph matching solver RRWM [2] to compute initial W .
- Compared against 3 baselines, matching accuracy was evaluated by the recall.

Dataset	Spectral	MatchLab	MatchALS	Ours-	Ours	Ours+
Hotel	0.51	0.64	0.58	0.63	0.69	1
House	0.14	0.79	0.75	0.79	0.90	1
Car	0.55	0.66	0.65	0.72	0.75	1
Duck	0.50	0.56	0.56	0.63	0.77	0.86
Face	0.62	0.93	0.94	0.95	0.99	1
Motorbike	0.25	0.28	0.27	0.40	0.61	1
Winebottle	0.64	0.71	0.72	0.75	0.82	1



BENCHMARK 2: DENSE SEMANTIC MATCHING

- Proposal flow [3] optimized by the proposed method.
- In proposal flow, correspondences of region proposals between images are transformed into a dense flow field.
- Proposal flow with selective search (SS), HOG descriptors and local-offset matching (LOM).
- 500 proposals extracted from each image. The proposed method (with $k=10$) treats each proposal as a **feature point**, where its center encodes the point coordinates.
- Dataset:** PF-WILLOW comprising 10 sub-classes.
- Metric:** % of corrected located keypoints (pixel distance below threshold), when transferring annotated keypoints from one image to another given the estimated flow.

Methods	car151	car150	car150	car150	car150	car150	car150	car150	car150	car150	car150
LOM+SS	0.89	0.82	0.76	0.70	0.69	0.71	0.78	0.91	0.72	0.77	0.87
LOM	0.86	0.76	0.77	0.69	0.68	0.78	0.76	0.89	0.77	0.80	0.85
DeepFlow	0.33	0.13	0.22	0.20	0.20	0.28	0.13	0.46	0.08	0.18	0.18
DeepF	0.44	0.20	0.34	0.27	0.33	0.12	0.10	0.40	0.17	0.18	0.18
SIFT flow	0.54	0.37	0.56	0.52	0.41	0.20	0.23	0.83	0.16	0.33	0.33
SRF	0.46	0.30	0.52	0.27	0.33	0.17	0.14	0.87	0.27	0.44	0.44
Zhou et al.	0.77	0.34	0.52	0.42	0.34	0.19	0.20	0.78	0.19	0.39	0.39



APPLICATION 2: AUTOMATIC LANDMARK ANNOTATION

- The proposed method was applied to the first 1000 images from the cat head dataset.
- Feature candidates were sampled from detected edges in images.
- # of selected features set to 10.
- Images (30) on the left show final selected features. Evidence correct correspondences across different instances (appearances and poses).
- Right column shows initial feature candidates (top) and manually annotated landmarks (bottom), for the first image.
- Notably, automatically selected features roughly coincide with human annotations.



CONCLUSIONS

- The proposed method solves the problem of semantic matching across multiple images.
- It establishes reliable feature correspondences among a collection of images satisfying both cycle consistency and geometric consistency.
- It outperforms previous multi-image matching methods.
- It is scalable to match thousands of images.

REFERENCES

- [1] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. In ICCV, 1992.
- [2] M. Cho, J. Lee, and K. M. Lee. Reweighted random walks for graph matching. In ICCV, 2010.
- [3] B. Han, M. Cho, C. Schmid, and J. Ponce. Proposal flow: Semantic correspondences from object proposals. In TPAMI, 2017.
- [4] X. Zhou, M. Zhu, and K. Danilidis. Multi-image matching via fast alternating minimization. In ICCV, 2015.

Better Poster



Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation

Min-Hung Chen¹ Baopu Li² Yingze Bao² Ghassan AlRegib¹ Zolt Kira¹

¹Georgia Institute of Technology ²Baidu USA

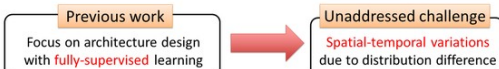
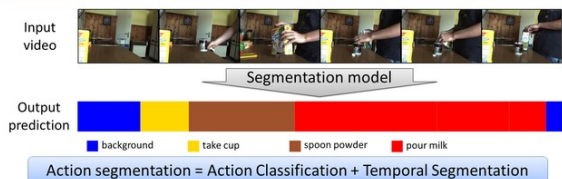


[Code] <https://github.com/cmhungsteve/SSTDA>
[Paper] <https://arxiv.org/abs/2003.02824>

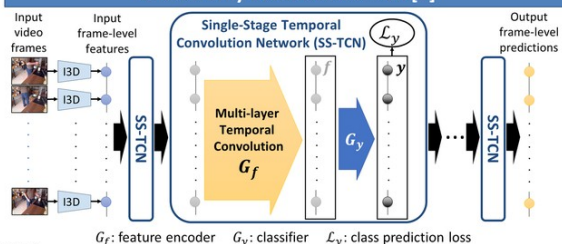
Summary

- Goal: address spatio-temporal variation problems using unlabeled videos
- Approach: **Self-Supervised Temporal Domain Adaptation (SSTDA)**
 - Multi-temporal domain prediction & adversarial domain confusion
 - Perform DA for multiple temporal scales
 - Learn feature representations with domain-invariant temporal dynamics
- Outperform other self-supervised methods and image-based DA methods
- Improve action segmentation by large margins using unlabeled target videos

Action Segmentation



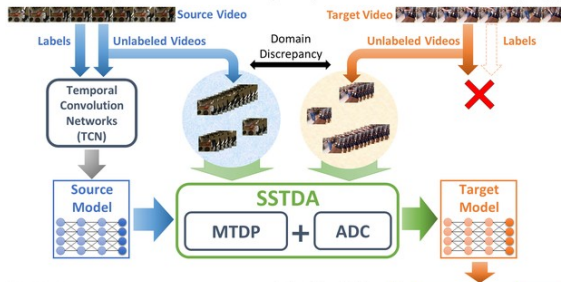
Source-only Baseline: MS-TCN [1]



[1] (CVPR 19)

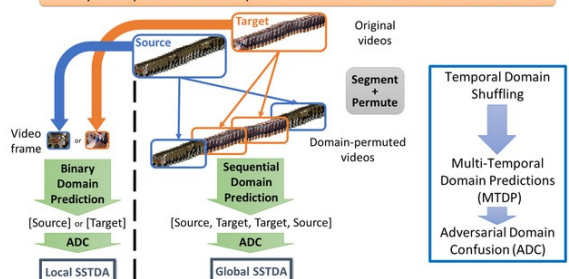
Main Idea

- Learn domain-invariant temporal dynamics using unlabeled videos
- Adopt fully-supervised methods (e.g. MS-TCN) to learn the source model
- Apply the proposed SSTDA to adapt the source model to target domains
 - SSTDA: reduce discrepancy across domains using unlabeled videos
 - Multi-Temporal Domain Predictions (MTDP) + Adversarial Domain Confusion (ADC) [2]



Self-Supervised Temporal Domain Adaptation (SSTDA)

binary & sequential domain predictions + adversarial domain confusion



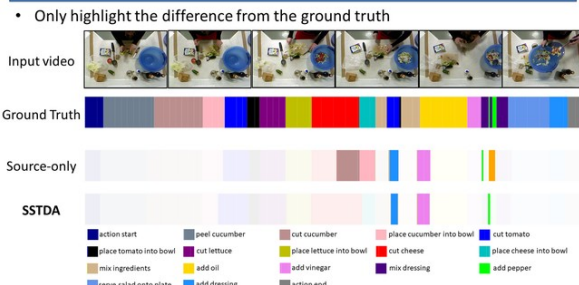
Quantitative Results

- Same backbone, MS-TCN, for all comparison
- Outperform within-domain self-supervised methods (e.g. VCOP [3]) and image-based DA methods (e.g. SWD [4])

Datasets	Approach	F1@10	F1@25	F1@50	Edit score	Accuracy
GTEA	Source-only†	86.5	83.7	71.9	81.3	76.5
	VCOP [3]	87.3	85.9	70.1	82.2	76.8
	SWD [4]	89.0	87.3	73.8	84.4	77.3
	SSTDA	90.0	89.1	78.0	86.2	79.8
50Salads	Source-only†	75.4	73.4	65.2	68.9	82.1
	VCOP [3]	75.8	73.8	65.9	68.4	82.3
	SWD [4]	78.2	76.2	67.4	71.6	81.9
	SSTDA	83.0	81.5	73.8	75.8	83.2
Breakfast	Source-only†	65.3	59.6	47.2	65.7	64.7
	VCOP [3]	68.5	62.9	50.1	67.9	66.7
	SWD [4]	68.6	63.2	50.6	69.1	67.1
	SSTDA	75.0	69.1	55.2	73.7	70.2

Jointly adapt domains with multiple temporal scales
→ Effectively reduce spatio-temporal variations for action segmentation

Qualitative Results (50Salads)



Better Poster



Feature Reconstruction-based Disentanglement for Pose-invariant Face Recognition

Xi Peng[†], Xiang Yu[†], Kihyuk Sohn[‡], Dimitris Metaxas[†], Manmohan Chandraker[§],
Rutgers, The State University of New Jersey[†], University of California, San Diego[‡], NEC Laboratories America[§]

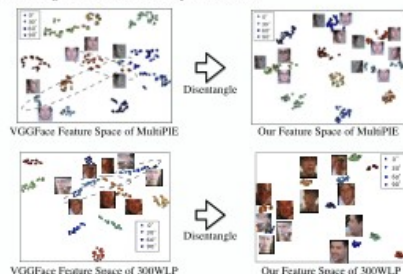
Project page: <https://sites.google.com/site/xipengcshomepage/iccy2017>



Highlights

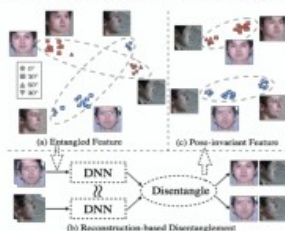
Challenge:

- ◆ Large pose variations are under-represented in face recognition datasets
- ◆ Face recognition features are not pose-invariant.



Our approach:

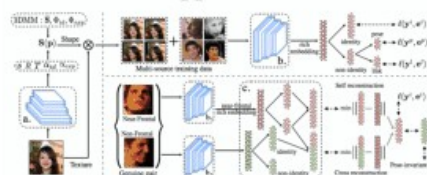
- ◆ Enhance pose diversity in training by generating non-frontal views.
- ◆ Feature reconstruction based metric learning to disentangle pose and identity



Main results:

- ◆ State-of-the-art results on controlled and uncontrolled datasets.
- ◆ Especially significant improvements for large poses.

Approach

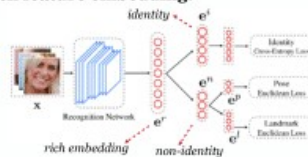


(a) Pose-variant face generation:

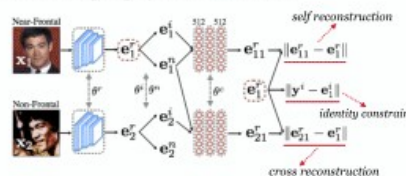


- ◆ Generate posed faces from frontal ones.
- ◆ Avoid artifacts caused by self occlusions
- ◆ Pose and landmark annotations for free.

(b) Rich feature embedding:



(c) Disentangling by reconstruction:



Results

	15	30	45	60	75	90	Avg
VGGFace	0.972	0.961	0.926	0.847	0.628	0.342	0.780
N-pair	0.990	0.983	0.971	0.944	0.811	0.468	0.861
GMA	1.000	1.000	0.904	0.952	0.725	0.550	0.838
MsVDN	1.000	0.991	0.921	0.897	0.810	0.706	0.887
Ours	0.972	0.966	0.956	0.927	0.857	0.749	0.905

	15	30	45	60	75	90	Avg
VGGFace	0.994	0.998	0.996	0.956	0.804	0.486	0.838
N-pair	1.000	0.996	0.993	0.962	0.845	0.542	0.899
Ours	1.000	0.999	0.995	0.994	0.978	0.940	0.988

Rank-1 recognition accuracy on MultiPIE (top) and 300WLP (bottom)

	Frontal-Frontal	Frontal-Profile	Verification accuracy on CFP dataset
FVE	98.67	91.97	
DR-GAN	97.84	93.41	
Human	96.24	94.57	
Ours	98.67	93.76	

Verification accuracy
on CFP dataset

Gallery and probe samples of MultiPTE (top) and 300WLP (bottom)



Some Failure cases in MultiPIE (left) and 300WLP (right)

[JOMA] Sharriss et al. "Generalized multiview analysis." In CVPR, 2018.

[VGGFace] Parkhi et al. "Deep face recognition." In BMVC, 2015.

[N-gin] Sohn et al. "Improved deep metric learning with multi-class n-gin loss objective." In NIPS, 2016.

[d-IN] Kan et al. "Multi-view deep network for cross-view classification." In CVPR, 2016.

[FVJ] Chen et al. "Fisher vector encoded deep neural features for unconstrained face verification." In IJCV, 2015.

[IR-GAN] Truu et al. "Disentangled representation learning for pose-invariant face recognition." In CVPR, 2017.

Better Poster



*Now at DeepMind

Convolutional neural network architecture for geometric matching

Ignacio Rocco^{1,2} Relja Arandjelović^{1,2,*} Josef Sivic^{1,2,3}

¹DI ENS, École normale supérieure, PSL Research University

²INRIA

³CIIRC, CTU in Prague

Goal

- Instance-level and category-level image alignment
- Output:** smooth dense correspondence field



Challenges

- Substantial appearance differences
- Presence of background clutter
- Lack of large annotated real image pair dataset

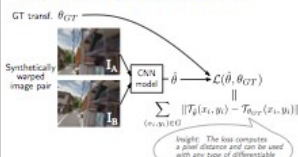
Contributions

- CNN architecture suitable for category-level alignment
- The model is trainable from synthetically warped image pairs
- Matching layer enables generalization to real image pairs

Overview

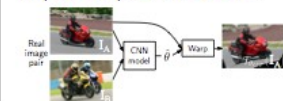
At training time:

- Inputs: Synthetically warped image pair
- Output:** Estimated parametric transformation $\hat{\theta}$



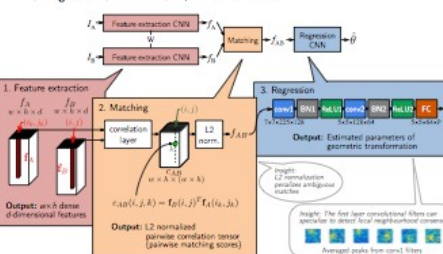
At evaluation time:

- Input: Real image pair
- Output:** Estimated parametric transformation $\hat{\theta}$



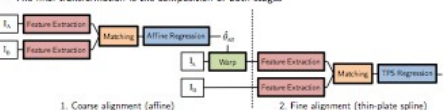
Model

- Three stage siamese CNN architecture mimicking the classical matching pipeline
- 1. Feature extraction CNN:** pre-trained VGG-16 model + per-column L2-normalization
- 2. Matching:** correlation layer + per-column L2-normalization
- 3. Regression CNN:** small CNN, trained from scratch



Coarse-to-fine matching architecture

- The same architecture can be applied with increasing geometric model complexity
- 1. Coarse alignment using an **affine transformation**
- 2. Refined alignment using a **thin-plate spline transformation**
- The final transformation is the composition of both stages



Training from synthetic imagery

- Training pairs:** generated by a real and a synthetically warped image



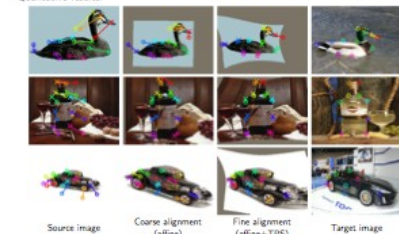
- Generalization:** We show that the method is relatively unaffected by the nature of the training images

Results on the Proposal Flow dataset

- Evaluated using annotated keypoints
- Metric: Percentage of correct keypoints (PCK)

Methods	PCK (%)
DeepFlow [1]	25
GMK [2]	27
SIFT Flow [3]	30
DSP [4]	29
Proposal Flow [5]	56
RANSAC with our features (affine)	47
Ours (affine)	49
Ours (affine + thin plate spline)	58
Ours (affine ensemble + thin plate spline)	57

Qualitative results:

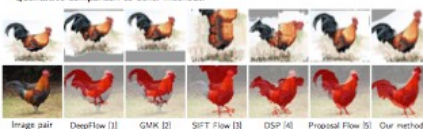


Results on the Caltech-101 dataset

- Evaluated using annotated object segmentation masks
- Metric: Label transfer accuracy (LT-ACC), Intersection over union (IoU), Localization error (LOC-ERR)

Methods	LT-ACC	IoU	LOC-ERR
DeepFlow [1]	0.74	0.40	0.34
GMK [2]	0.77	0.42	0.34
SIFT Flow [3]	0.75	0.48	0.32
DSP [4]	0.77	0.47	0.35
Proposal Flow [5]	0.78	0.50	0.25
Ours (affine)	0.79	0.51	0.25
Ours (affine + thin-plate spline)	0.82	0.58	0.25

Qualitative comparison to other methods:



References

- [1] P. Weinzaedl, et al. DeepFlow: Large displacement optical flow with deep matching. In Proc. ICCV, 2013
- [2] O. Duchenne, et al. A graph-matching kernel for object categorization. In Proc. ICCV, 2011
- [3] C. Liu, et al. SIFT Flow: Dense correspondence across scenes and its applications. IEEE PAMI, 2011
- [4] J. Kim, et al. Deformable spatial pyramid pooling for fast dense correspondences. In Proc. CVPR, 2013
- [5] B. Han, M. Cho, C. Schmid, and J. Ponce. Proposal Flow. In Proc. CVPR, 2016