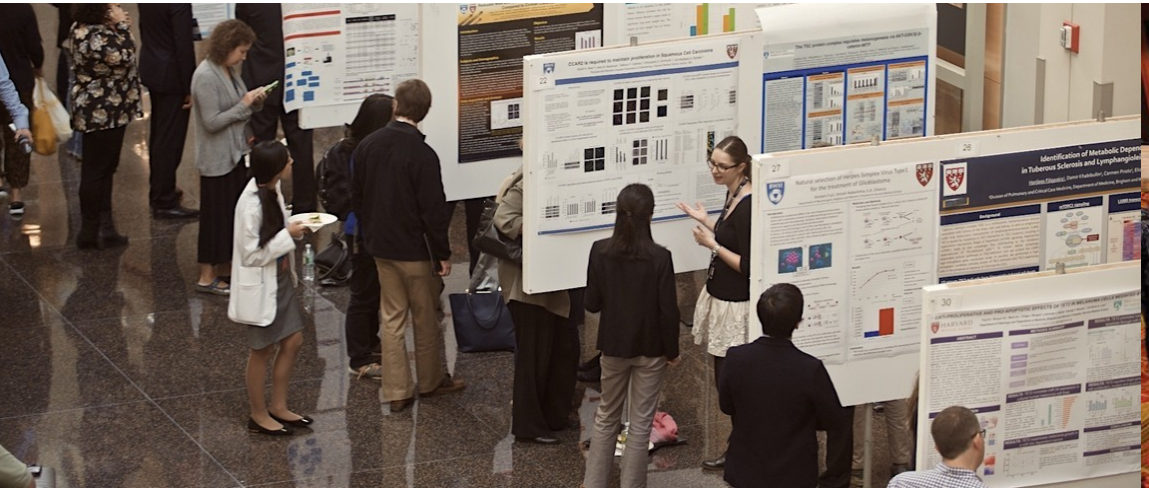
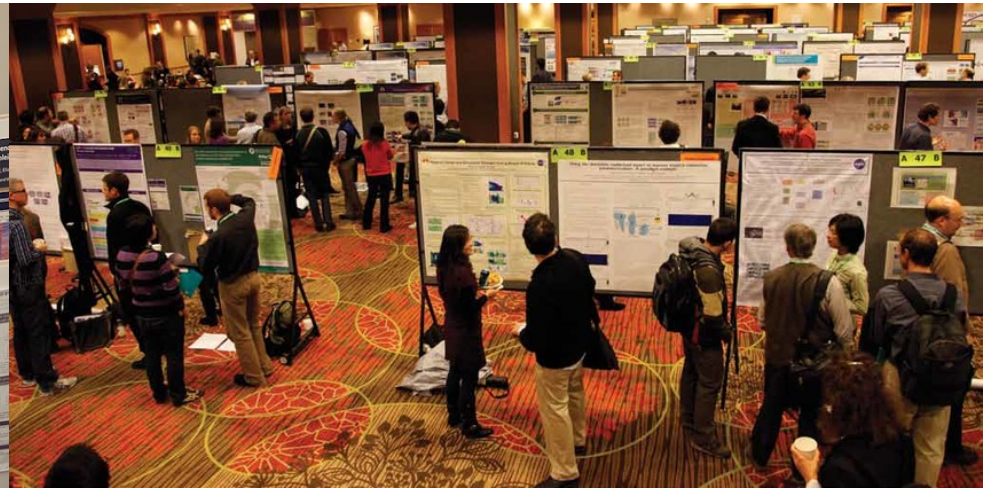
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

3D Data Processing
Poster Session

Alberto Pretto

Poster Session





Poster Session

- **Groups**: 2 students
- **When**: June 16th 2023, h.10:30 – 12:30
- **Where**: in the corridor in front of the exit door of Ve classroom (third floor of DEI/G)

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 preocessing/semantic segmentation
  - 3D object pose estimation
  - 3D Shape registration
  - 3D sensors
  - -...
- 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 attenders.

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 or LibreOffice Impress.

Poster Submission

- Through the e-learning website (a submission link will be shared soon) **by June 14th h. 24**. I will print the posters you submitted, so the day of presentation I'll bring the posters
- You may also bring your poster on June 16th, 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 Daniilidis
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 $1 < i < n$ 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 $\boldsymbol{P}_{ij} \in \{0,1\}^{p_i \times p_j}$.
- **Pairwise matching**: Individual pairs $\boldsymbol{P}_{ij}$ can be estimated using the Hungarian algorithm or approximated graph matching algorithms. Said estimates will be denoted by $\boldsymbol{P}_{ij} \approx \boldsymbol{W}_{ij} \in \mathbb{R}^{p_i \times p_j}$.
- **Cycle consistency**: Constraint used in the multi-image case. For any triplet (i,j,z), it must hold: $\boldsymbol{P}_{ij} = \boldsymbol{P}_{iz}\,\boldsymbol{P}_{zj}$. Given, $\boldsymbol{X}_i \in \{0,1\}^{p_i \times u}$, the correspondence between image i and the universe of unique features in the collection. The set $\{\boldsymbol{P}_{ij} | \forall i,j\}$ is cyclically consistent iff $\boldsymbol{P}$ can be factorized as $\boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X} \in \{0,1\}^{mn \times u}$ and:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{11} & \cdots & \boldsymbol{P}_{1n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{P}_{n1} & \cdots & \boldsymbol{P}_{nn} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}$$

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 $\boldsymbol{X}_i$ through minimization subject to a **sparsity constraint** dependent on k: $\min_{\boldsymbol{X}} \frac{1}{4}\left\|\boldsymbol{W} - \boldsymbol{X}\boldsymbol{X}^T\right\|_F^2,\ s.t.\ \boldsymbol{X}_i \in \{0,1\}^{p_i \times k}$.
- Where, $\boldsymbol{W} \in \mathbb{R}^{m \times m}$ is the concatenation of $\boldsymbol{W}_{ij}$, i.e. the initial (non-consistent) estimation of $\boldsymbol{P}$ computed by **pairwise solvers**.
- **Geometric constraint**: Spatial relationships (ex. nose below eyes). Low **rank constraint** on the measurement matrix $\hat{\boldsymbol{M}}$ 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. $\hat{\boldsymbol{M}}$ still approximated by a low-rank (r) matrix.
- $\hat{\boldsymbol{M}} \in \mathbb{R}^{2n \times k}$ built with the coordinates of the k feature points selected from each image $\hat{\boldsymbol{M}}_i = \boldsymbol{C}_i \boldsymbol{X}_i$, where $\boldsymbol{C}_i \in \mathbb{R}^{2 \times p_i}$.
- Geometric constraint imposed by minimizing: $\min_{\boldsymbol{X},\boldsymbol{Z}} \frac{1}{2}\sum_{i=1}^{n} \|\boldsymbol{C}_i\boldsymbol{X}_i - \boldsymbol{Z}_i\|_F^2,\ s.t.\ rank(\boldsymbol{Z}) \le r$, where $\boldsymbol{Z} \in \mathbb{R}^{2n \times k}$ is an auxiliary variable.

FORMULATION

- Combining cycle and geometric consistency terms gives the final optimization problem:

$$\min_{\boldsymbol{X},\boldsymbol{Z}} \frac{1}{4}\left\|\boldsymbol{W} - \boldsymbol{X}\boldsymbol{X}^T\right\|_F^2 + \frac{\lambda}{2}\sum_{i=1}^{n} \|\boldsymbol{C}_i\boldsymbol{X}_i - \boldsymbol{Z}_i\|_F^2$$
$$s.t.\ \boldsymbol{X}_i \in \{0,1\}^{p_i \times k} \ \wedge\ rank(\boldsymbol{Z}) \le r$$

- Where λ controls the impact of the geometric constraint.

PRACTICAL IMPLEMENTATION

- Replace $\boldsymbol{X}$ in the first term with an auxiliary variable $\boldsymbol{Y} \in \mathbb{R}^{m \times k}$, i.e. real matrix representing a permutation matrix.
- And add a soft constraint to push $\boldsymbol{Y}$ towards $\boldsymbol{X}$.

$$\min_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}} \frac{1}{4}\left\|\boldsymbol{W} - \boldsymbol{Y}\boldsymbol{Y}^T\right\|_F^2 + \frac{\lambda}{2}\sum_{i=1}^{n} \|\boldsymbol{C}_i\boldsymbol{X}_i - \boldsymbol{Z}_i\|_F^2 + \frac{\rho}{2}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$
$$s.t.\ \boldsymbol{X}_i \in \{0,1\}^{p_i \times k} \ \wedge\ rank(\boldsymbol{Z}) \le r \wedge \boldsymbol{Y} \in \mathcal{C}$$
$$Where, \mathcal{C}: \mathbf{0} \le \boldsymbol{Y} \le \mathbf{1}, \mathbf{0} \le \boldsymbol{Y}_i\mathbf{1} \le \mathbf{1}, \boldsymbol{Y}_i^T = \mathbf{1}$$

- Apply block coordinate descent, i.e., alternately updating one variable $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ while fixing the others.
  - $\boldsymbol{Y}$ updated via projected gradient descent.
  - Each $\boldsymbol{X}_i$ updated via the Hungarian algorithm.
  - $\boldsymbol{Z}$ updated via singular value decomposition.
- Empirically set λ as 1, r as 4, ρ as sequence $(1,10,100)$.
- Reliably initialize $\boldsymbol{Y}$ by ignoring geometric constraint, $\min_{\boldsymbol{Y}} \frac{1}{4}\left\|\boldsymbol{W} - \boldsymbol{Y}\boldsymbol{Y}^T\right\|_{F'}^2\ s.t. \boldsymbol{Y} \in \mathcal{C}$. Discretize $\boldsymbol{Y}$ to init $\boldsymbol{X}$.

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, $\boldsymbol{W}$ obtained from the Hungarian algorithm. Whereas, Ours+ employed the graph matching solver RRWM [2] to compute initial $\boldsymbol{W}$.
- Compared against 3 baselines, matching accuracy was evaluated by the recall.

Dataset	Spectral	MatchLift	MatchALS	Ours-	Ours	Ours+
Hotel	0.53	0.64	0.58	0.63	0.90	1
House	0.74	0.79	0.75	0.79	0.93	1
Car	0.55	0.66	0.65	0.72	0.75	1
Duck	0.59	0.56	0.56	0.63	0.77	0.88
Face	0.92	0.93	0.94	0.95	0.95	1
Motorbike	0.25	0.28	0.27	0.40	0.61	1
Winebottle	0.64	0.71	0.72	0.73	0.82	1

- Next figure shows results with (bottom) and without (top) geometric constraint. True and false matches are shown in blue and red, respectively.

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	car(S)	car(G)	car(M)	duc(S)	mot(S)	mot(G)	mot(M)	win(w/o C)	win(w/C)	win(M)
LOM + Ours	0.89	0.62	0.56	0.70	0.49	0.31	0.28	0.91	0.52	0.72
LOM	0.86	0.58	0.52	0.65	0.48	0.28	0.28	0.91	0.37	0.65
DeepFlow	0.33	0.13	0.22	0.20	0.20	0.08	0.13	0.46	0.08	0.18
GMK	0.48	0.25	0.34	0.27	0.31	0.12	0.15	0.41	0.17	0.18
SIFT Flow	0.54	0.37	0.36	0.32	0.41	0.20	0.23	0.85	0.16	0.33
DSP	0.46	0.30	0.32	0.25	0.31	0.15	0.14	0.85	0.25	0.64
Zhou et al.	0.77	0.56	0.52	0.42	0.34	0.19	0.20	0.78	0.19	0.38

- Next figure shows source image warped to target image using dense correspondences from proposal flow, and correspondences optimized by the proposed method.

APP. 1: OBJECT-CLASS MODEL RECONSTRUCTION

- **Goal**: Reconstruction of objects from images of different object instances.
- **Previous work**: rely on annotated keypoints in images, or on object masks to remove background [4].
- **Proof of concept**: this experiment demonstrates that the proposed method produces consistent correspondences for reconstruction without using any manual annotation, and it **doesn't need object masks** thanks to its capability to **prune nonrepeatable features** in the background.
- Uniformly sample feature candidates on image edges detected by the structured forests. Deep features used as descriptors and RRWM solver for initial matching.
- For reconstruction, simply run affine reconstruction through the factorization method [1].
- **Datasets**: FG3DCar dataset (37 left-view sedan images), and 30 images of different motorbikes with similar views.
- **Results**: shown below. Reconstruction (bottom, two viewpoints) from just 4 images.

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 satisficing 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 IJCV, 1992.
[2] M. Cho, J. Lee, and K. M. Lee. Reweighted random walks for graph matching. In ECCV, 2010.
[3] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow: Semantic correspondences from object proposals. In T-PAMI, 2017.
[4] X. Zhou, M. Zhu, and K. Daniilidis. Multi-image matching via fast alternating minimization. In ICCV, 2015.

Better Poster

Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation

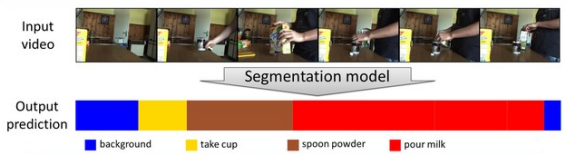
Min-Hung Chen[1] Baopu Li[2] Yingze Bao[2] Ghassan AlRegib[1] Zsolt Kira[1]
[1]Georgia Institute of Technology [2]Baidu USA

[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

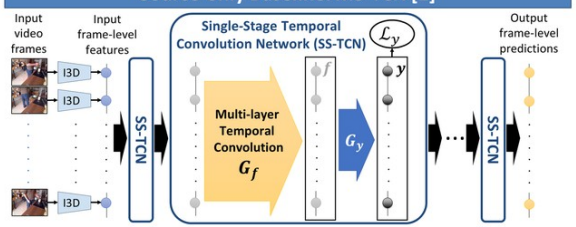


Action segmentation = Action Classification + Temporal Segmentation

Previous work: Focus on architecture design with fully-supervised learning

Unaddressed challenge: Spatial-temporal variations due to distribution difference

Source-only Baseline: MS-TCN [1]

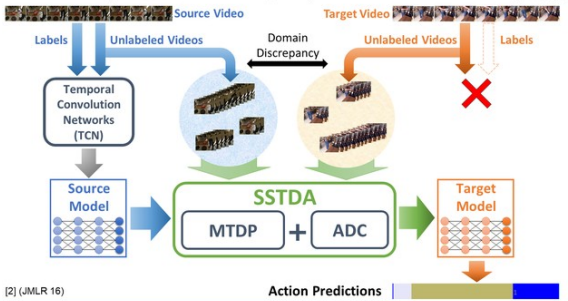


G_f: feature encoder G_y: classifier $\mathcal{L}_y$: class prediction loss

[1] (CVPR 19)

Main Idea

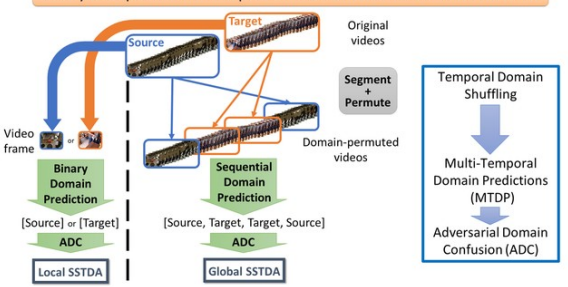
- Learn domain-invariant temporal dynamics using unlabeled videos
- Adopt fully-supervised methods (e.g. MS-TCN) to learn the source model
- Apply the proposed *SSTDA* to adapt the source model to target domains
  - *SSTDA*: reduce discrepancy across domains using unlabeled videos
  - Multi-Temporal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



[2] (JMLR 16)

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])

[3] (CVPR 19), [4] (CVPR 19)

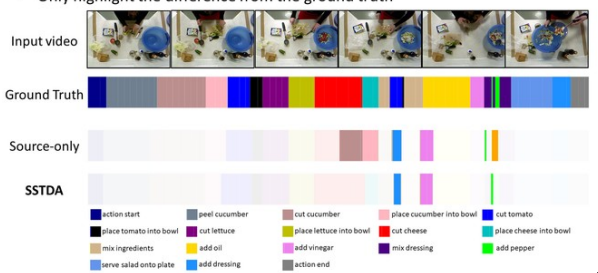
*Source-only: MS-TCN w/o target data

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)

- Only highlight the difference from the ground truth



Better Poster

NEC Laboratories America

ICCV'17

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 ‡

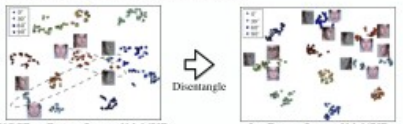
Project page: https://sites.google.com/site/xipengcshomepage/iccv2017

RUTGERS

Highlights

Challenge:

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

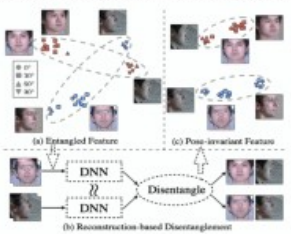


VGGFace Feature Space of MultiPIE — Our Feature Space of MultiPIE

VGGFace Feature Space of 300WLP — Our Feature Space of 300WLP

Our approach:

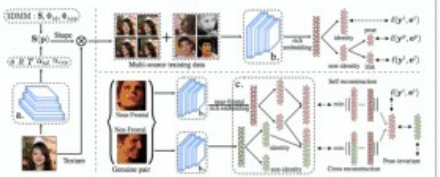
- Enhance pose diversity in training by generating non-frontal views.
- Feature reconstruction based metric learning to disentangle pose and identity.



Main results:

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

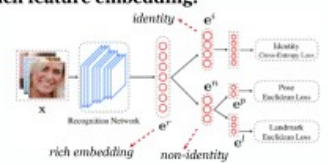
Approach



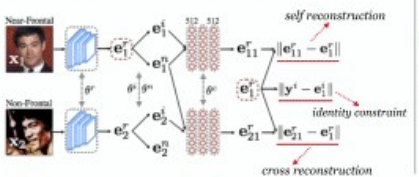
(a) Pose-variant face generation:

- 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	0.852	0.725	0.550	0.838
MvDN	**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	0.859
Ours	**1.000**	**0.999**	0.995	**0.994**	**0.978**	**0.940**	**0.980**

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

	Frontal-Frontal	Frontal-Profile
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 MultiPIE (top) and 300WLP (bottom).

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

[GMA] Sharma et al. "Generalized multiview analysis." In CVPR, 2012.
[VGGFace] Parkhi et al. "Deep face recognition." In BMVC, 2015.
[N-pair] Sohn et al. "Improved deep metric learning with multi-class n-pair loss objective." In NIPS, 2016.
[MvDN] Kan et al. "Multi-view deep network for cross-view classification." In CVPR, 2016.
[FVE] Chen et al. "Fisher vector encoded deep neural features for unconstrained face verification." In ICIP, 2016
[DR-GAN] Tran et al. "Disentangled representation learning gan for pose-invariant face recognition." In CVPR, 2017.

Better Poster