facebook
Artificial Intelligence Research

August 29th, 2019

Geometry in Computer Vision

Natalia Neverova
Research Scientist, Facebook AI

I. Learning correspondences

II. 3D reconstruction

III. Generation

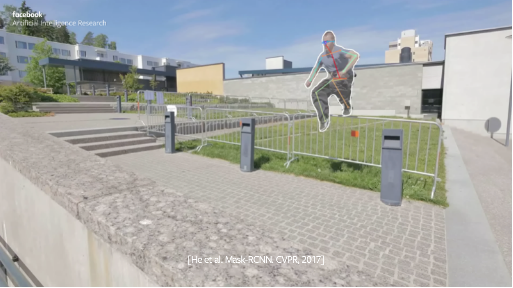
Application: human-centered tasks

3D-fying panoptic perception in the wild

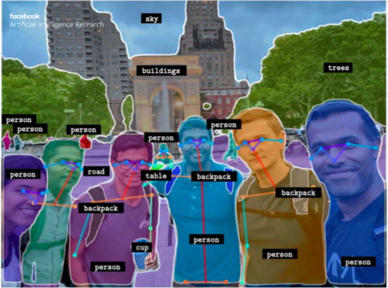
Object boxes

Masks

[Slide credit Jitendra Malik]



[He et al. Mask-RCNN. CVPR, 2017]

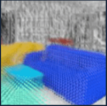

facebook
Artificial Intelligence Research
sky
buildings
trees
person
person
person
person
person
person
road
table
backpack
person
backpack
backpack
cup
person
person
person
person

3D-fying panoptic perception in the wild

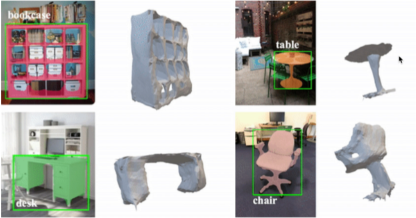
Object boxes

Masks

3D geometry

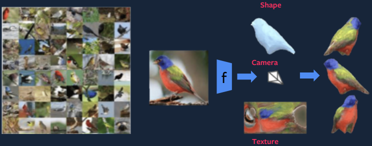


[Slide credit Jitendra Malik]



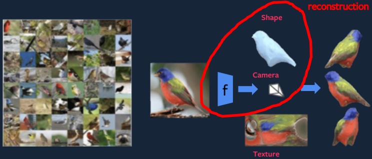
[Gkioxari et al. Mesh-RCNN. ICCV, 2019]

Objects: shape and appearance decomposition



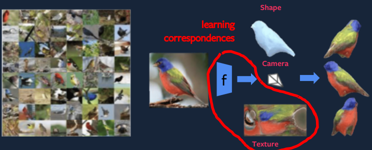
[Kanazawa et al. Learning Category-Specific Mesh Reconstruction from Image Collections. ECCV, 2018]

Objects: shape and appearance decomposition



[Kanazawa et al. Learning Category-Specific Mesh Reconstruction from Image Collections. ECCV, 2018]

Objects: shape and appearance decomposition



[Kanazawa et al. Learning Category-Specific Mesh Reconstruction from Image Collections. ECCV, 2018]

I. Learning correspondences

Learning correspondences: image to image

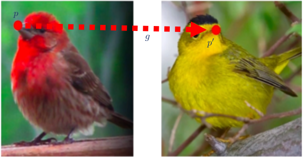
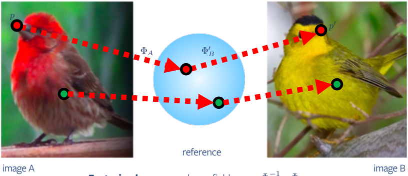


image A

image B

Let g be the **correspondence field** between images A and B

Learning correspondences



Factorized correspondence field: $g = \Phi_B^{-1} \circ \Phi_A$

3D geometry is **irrelevant,** we only need **an index set** over the object surface

Learning correspondences: image to 3D model

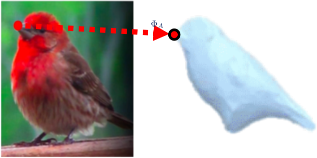


image A

3D model

Learning correspondences

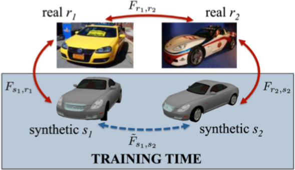
Supervised

- model-driven;
- data-driven.

Un/self-supervised

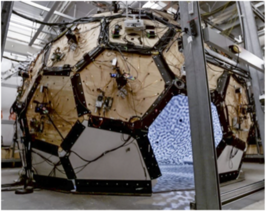
- equivariance;
- cycle consistency.

Model-driven: synthetic data



[Zhou et al. Learning Dense Correspondence via 3D-guided Cycle Consistency. ECCV, 2016]

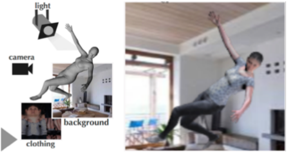
Synthetic data: articulated objects?



[Joo et al. Panoptic studio: A massively multiview system for social interaction capture. PAMI, 2016]

[Loper, Mahmood, Romero, Pons-Moll, Black.
SMPL: A Skinned Multi-Person Linear Model. SIGGRAPH Asia, 2015]

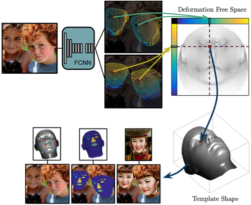
Synthetic data: articulated objects?



[Varol et al. "Learning from Synthetic Humans". CVPR, 2017]

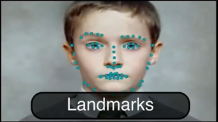
Very different image statistics!

Model-driven: sparse annotations + fitting

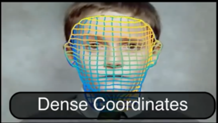


[Güler et al. “DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild”. CVPR, 2017]

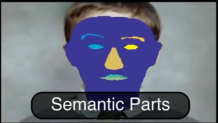
Input Image



Landmarks

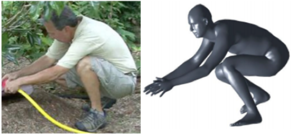


Dense Coordinates



Semantic Parts

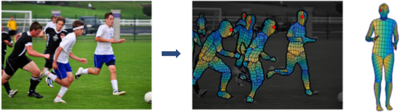
Full body articulation?



[Lassner et al. “Unite the People: Closing the Loop Between 3D and 2D Human Representations”. CVPR, 2017]

Poor approximation of real data!

Data-driven approach: DensePose



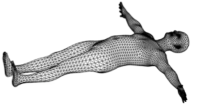
[Güler et al. “DensePose: learning dense correspondences in the wild”. CVPR, 2018]

Eliminates dependency on a **specific 3D model** and its **expressivity** (as long as **semantics is preserved**)

No domain gap, annotations are **easier to obtain**

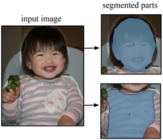
Human annotation **errors can be significant** due too ambiguities

Dense correspondence task

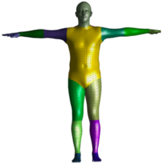


COCO-DensePose dataset

Annotation task 1: body part segmentation

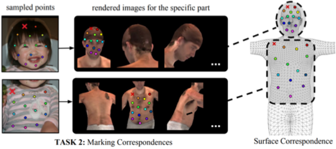


TASK 1: Part Segmentation

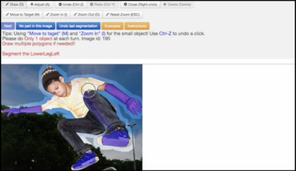


COCO-DensePose dataset

Annotation task 2: marking sparse correspondences


sampled points
rendered images for the specific part
TASK 2: Marking Correspondences
Surface Correspondence

COCO DensePose:
Collecting Data



Tips: Using "Move to taget" (M) and "Zoom In" (I) for the small object! Use Ctrl-Z to undo a click.
Please do Only 1 object at each turn. Image id: 195
Draw multiple polygons if needed!!

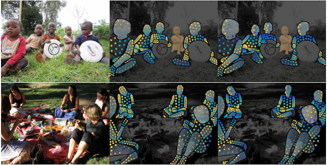
Segment the LowerLegLeft

Task - 1

Part Segmentation

COCO-DensePose dataset

50 annotated instances, 5 million correspondences (~100 points/image)



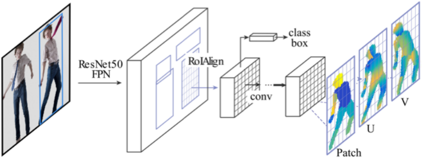
Evaluation metric: geodesic distance

For instance based frameworks:

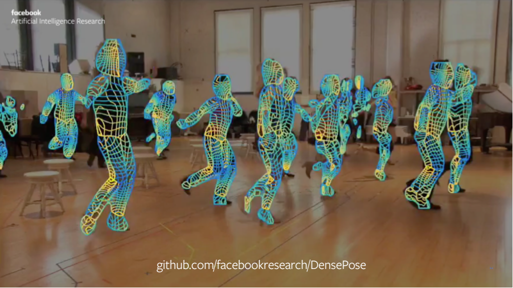
$$\mathrm{GPS}_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2\kappa^2} \right)$$

geodesic point similarity (GPS)

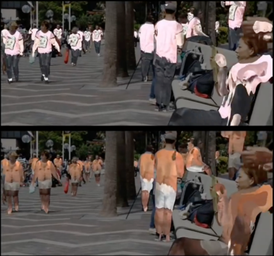
Architecture: DensePose-RCNN

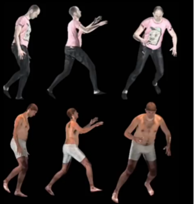


[He, Gkioxari, Dollar, Girshink. Mask-RCNN. ICCV, 2017]



github.com/facebookresearch/DensePose

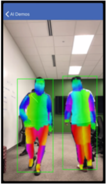




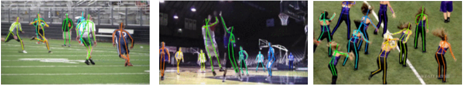
Textures taken from SURREAL dataset.

Varol, Gül, et al. "Learning from synthetic humans." CVPR 2017.

Real-time demos on desktop & mobile

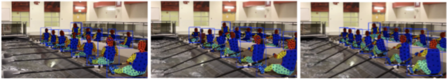


DensePoseTrack dataset



[Andriluka et al. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. CVPR, 2018]

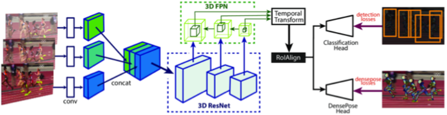
DensePoseTrack dataset



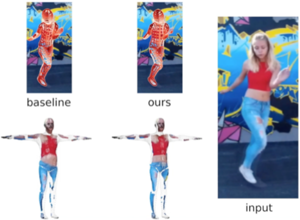
[Neverova et al. Slim DensePose: Thrifty Learning with Motion Cues. CVPR, 2019]

Labeled images: 1680 / 782 (training / validation)
Instances: 8274 / 4753
Correspondences: 800142 / 459348
Every 2nd frame for 4 frames, every 8th frame otherwise
Ignored: instances with <6 keypoints, severe motion blur

Flow-guided 3D DensePose-RCNN



[Khalidov et al., 2019]



[Neverova et al. Slim DensePose: Thrifty Learning with Motion Cues. CVPR, 2019]

Learning correspondences

Supervised

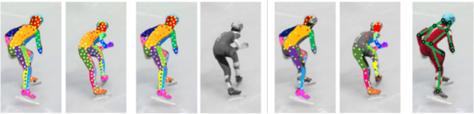
- model-driven;
- data-driven.

Un/self-supervised

- equivariance;
- cycle consistency.

Learning with less supervision?

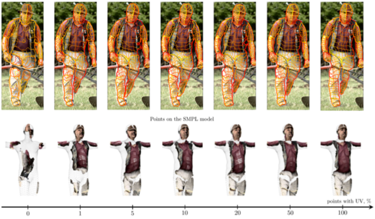
Cost efficient annotation process



full annotations full dense annotations sparse annotations keypoints

[Neverova et al. Slim DensePose: Thrifty Learning with Motion Cues. CVPR, 2019]

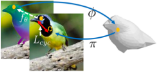
Cost efficient annotation process



Correspondences by self-supervision

GT propagation & equivariance

[Neverova et al. Slim DensePose: Thrifty Learning with Motion Cues. CVPR, 2019]

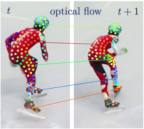


cycle-consistency

[Kulkarni et al. Canonical Surface Mapping via Geometric Cycle Consistency. ICCV, 2019]

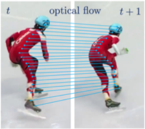
GT propagation vs equivariance

Transfer a given label to a new frame



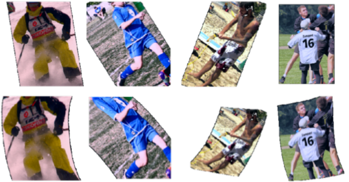
GT propagation

Constrain unknown labels to be consistent

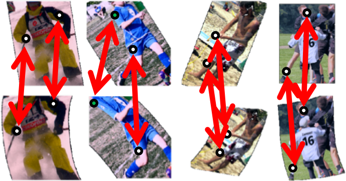


equivariance

Synthetic equivariance: thin-plane splines (TPS)

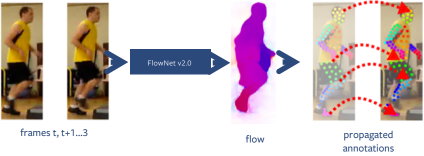


Synthetic equivariance: thin-plane splines (TPS)



The **known mapping between points in a pair of original-deformed frames** is used both for **data augmentation (sparsely)** and enforcing **equivariance (densely).**

Flow-guided temporal equivariance



frames t, t+1...3

FlowNet v2.0

flow

propagated annotations

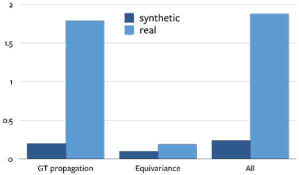
Optical flow is used both for **data augmentation (sparse points)** and enforcing **inter-frame temporal equivariance (densely)**

Flow-guided temporal equivariance

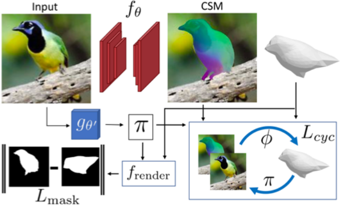
Real transform >> synthetic

GT propagation >> equivariance

Combination > individual



Cycle-consistency



facebook
Artificial Intelligence Research

II. Reconstruction

3D reconstruction

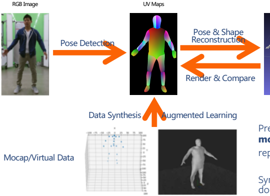
3D-supervised

- synthetic data;
- multi-view data / video;
- manual annotations (?).

2D-supervised

Model-based 3D reconstruction

facebook
Reality Labs



Predicting **parameters of the SMPL model** based on **DensePose** representation rather than RGB

Synthetic data helps due to smaller domain gap in this space

[DenseRac: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. ECCV, 2018]

Synthetic Data from Virtual World

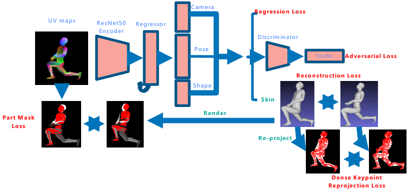
facebook
Reality Labs

Mixamo
(www.mixamo.com)

Offering free animated 3D characters

Thousands of customizable 3D animations

Differentiable rendering



UV maps

ResNet50 Encoder

Regressor

Camera

Pose

Shape

Regression Loss

Discriminator

Yes/No

Adversarial Loss

Reconstruction Loss

Skin

Render

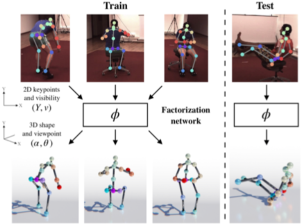
Re-project

Part Mask Loss

Dense Keypoint Reprojection Loss

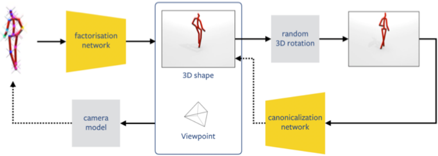
Learning with less supervision?

3D reconstruction with 2D supervision

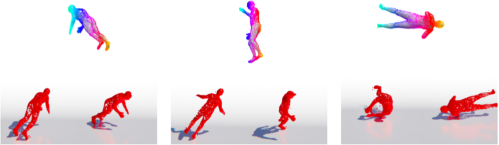


[Novotny, Ravi et al. C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion. ICCV, 2019]

Canonicalization network



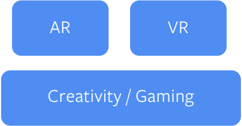
From sparse landmarks to dense annotations



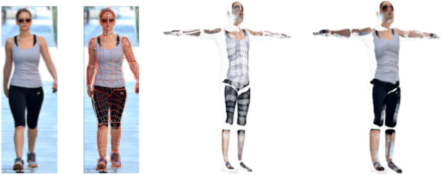
Qualitative results on **synthetic renderings** using the SMPL model

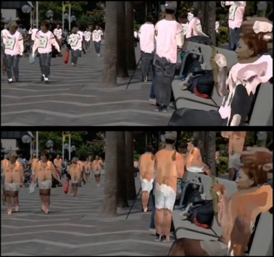
III. Generation

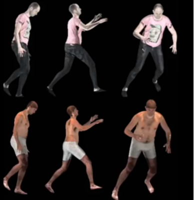
Generation



Mapping from image to texture space

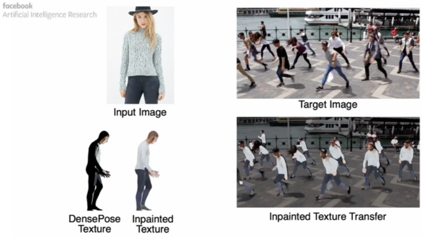






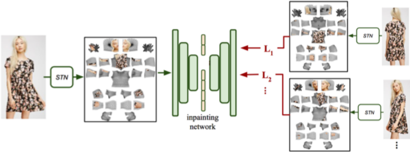
Textures taken from SURREAL dataset.

Varol, Gül, et al. "Learning from synthetic humans." CVPR 2017.



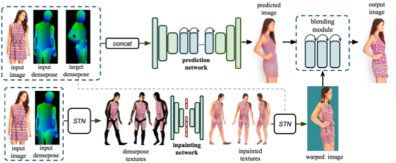
[Neverova et al. Dense Pose Transfer [CVPR, 2019]

Texture inpainting in UV space



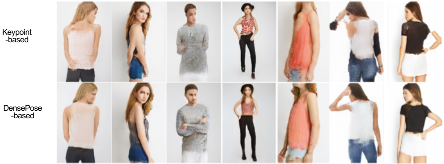
The inpainting network learns to reconstruct **full body texture** from **partial observations** by autoencoding in a normalized texture space

Two stream model



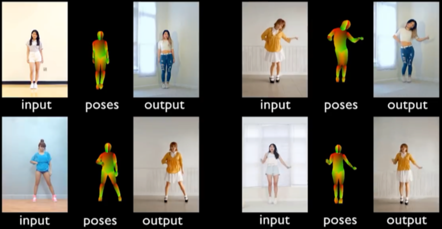
The inpainting network introduces **generalization over the pose** space for free

DensePose vs sparse keypoints conditioning



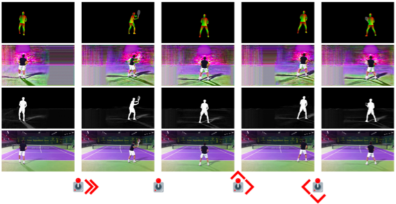
Conditioning on DensePose resolves **ambiguity in z-ordering** and encourages **anatomical plausibility**

From fits-them-all to personalized models



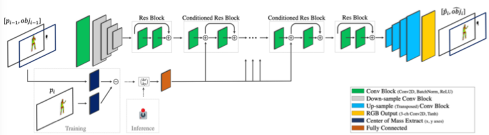
[Wang, Liu, Zhu, Liu, Tao, Kautz, Catanzaro. Video-to-Video Synthesis. NeurIPS, 2018]

Vid2game: creating controllable characters



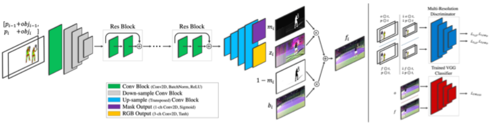
[Gafni, Wolf, Taigman. Vid2Game: Controllable Characters Extracted from Real-World Videos. arXiv:1904.08379, 2019]

Vid2game: creating controllable characters


$[p_{i-1}, obj_{i-1}]$
Res Block
Conditioned Res Block
Conditioned Res Block
Res Block
$[\hat{p}_i, \widehat{obj}_i]$
p_i
$\binom{\Delta x}{\Delta y}$
Training
Inference
🟩 Conv Block (Conv2D, BatchNorm, ReLU)
▨ Down-sample Conv Block
🟦 Up-sample (Transposed) Conv Block
🟨 RGB Output (3-ch Conv2D, Tanh)
■ Center of Mass Extract (x, y axes)
🟧 Fully Connected

DensePose representation of the next frame is predicted conditioned on a current DensePose and an instruction

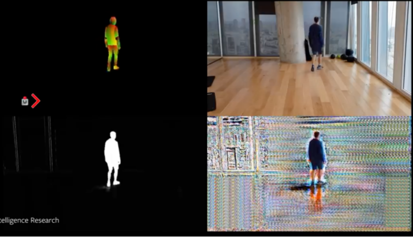
Vid2game: creating controllable characters



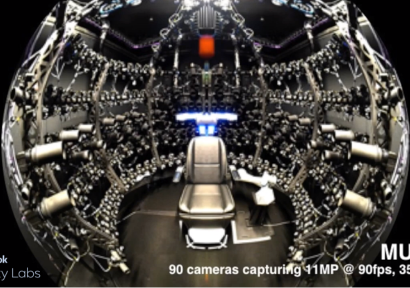
The new frame is rendered through decomposition: **active character** (predicted) + **background** (copied)

Target Video

Evaluation - Walking (Controllable Results)



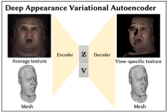
… and more personalization



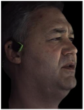
MUGSY

90 cameras capturing 11MP @ 90fps, 350 lights

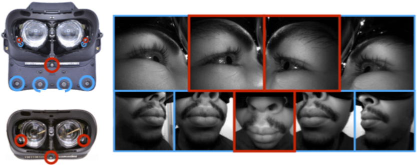
facebook
Reality Labs

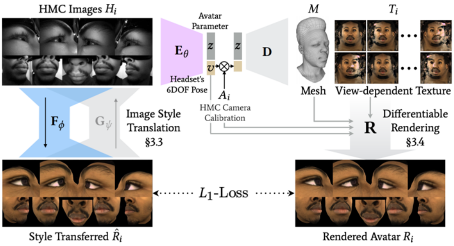


Ground Truth
(novel view)

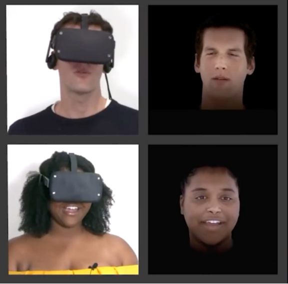


Rendered Avatar
(novel view)




HMC Images H_i
Avatar Parameter
$\mathbf{E}_\theta$
z
v
Headset's 6DOF Pose
A_i
z
$\mathbf{D}$
M
Mesh
T_i
View-dependent Texture
$\mathbf{F}_\phi$
$\mathbf{G}_\psi$
Image Style Translation §3.3
HMC Camera Calibration
$\mathbf{R}$
Differentiable Rendering §3.4
L_1-Loss
Style Transferred $\hat{R}_i$
Rendered Avatar R_i

facebook
Reality Labs



Questions?

facebook
Artificial Intelligence Research