Learning the Depths of Moving People by Watching Frozen People

Zheng Li, Tali Dekel, Forrester Cole,
Richard Tucker, Noah Snavely, Ce Liu,
William T. Freeman
Google Research

Presenters: Kiranpreet Kaur, Weigang Yi, An-Hsiu Cheng

Motivation

It’s easy for human beings to predict the 3D outline of an object by 2D images where it doesn’t matter if the object is static or moving

Static

Moving

Motivation (Cont.)

However, it’s not easy for computers to do that even if a object is static.

1. The existing methods for estimating depth are only meant for static objects. Moving objects violate the epipolar constraint used in 3D vision.

Moreover, **moving objects are often treated as noise** or **outliers** in existing Structure-from- Motion (SfM) and Multi-view Stereo (MVS) methods.

Motivation (Cont.)

However, it’s not easy for computers to do that even if a object is static.

1. The existing methods for estimating depth are only meant for static objects. Moving objects violate the epipolar constraint used in 3D vision.

Moreover, **moving objects are often treated as noise** or outliers in existing Structure-from- Motion (SfM) and Multi-view Stereo (MVS) methods.

2. The existing methods only estimate depth on a single RGB image.

Motivation (Cont.)

However, it’s not easy for computers to do that even if a object is static.

1. The existing methods for estimating depth are only meant for static objects. Moving objects violate the epipolar constraint used in 3D vision.

Moreover, **moving objects are often treated as noise** or outliers in existing Structure-from- Motion (SfM) and Multi-view Stereo (MVS) methods.

2. The existing methods only estimate depth on a single RGB image.

3. Depth sensors, such as Kinect, can provide useful data but it is limited to indoor environments and requires significant manual work.

Problem Statement

Focus: predicting accurate dense depth of humans from videos using a data-driven approach.

- where both the camera and people in the scene are naturally moving.
- creating a dataset that can be used by other models.

Problem Statement

Focus: predicting accurate dense depth of humans from videos using a data-driven approach.

- where both the camera and people in the scene are naturally moving.
- creating a dataset that can be used by other models.

Reasons:

- In augmented reality (AR), humans are important objects in the scene.
- Human motion is articulated and difficult to model.

RW: **Learning-based depth prediction**

- Predicting dense depth from a single RGB image.
- Some methods also consider multiple images

Single Image Depth Perception in the wild

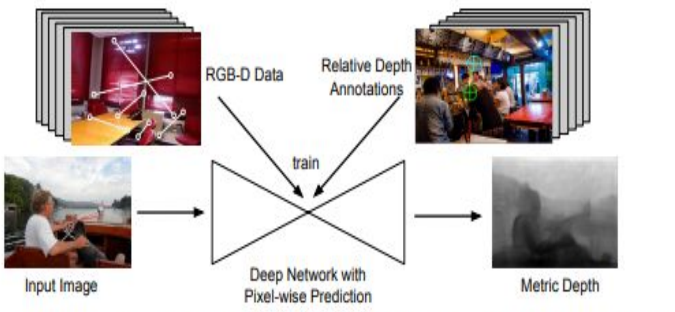
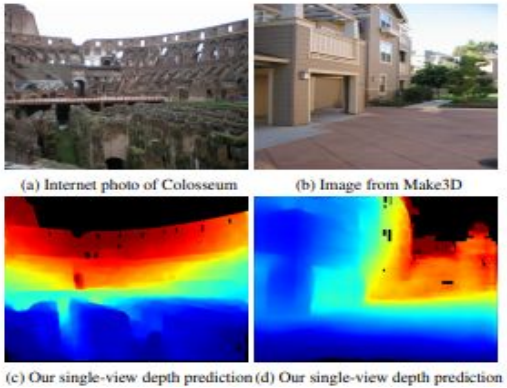

RGB-D Data
Relative Depth Annotations
train
Input Image
Deep Network with Pixel-wise Prediction
Metric Depth

Figure 1: We crowdsource annotations of relative depth and train a deep network to recover depth from a single image taken in unconstrained settings ("in the wild").

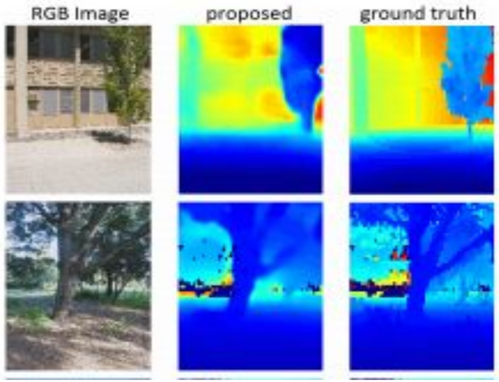
MegaDepth: Learning Single-View Depth Prediction from Internet Photos



(a) Internet photo of Colosseum (b) Image from Make3D

(c) Our single-view depth prediction (d) Our single-view depth prediction

Deeper Depth Prediction with Fully Convolutional Residual Networks



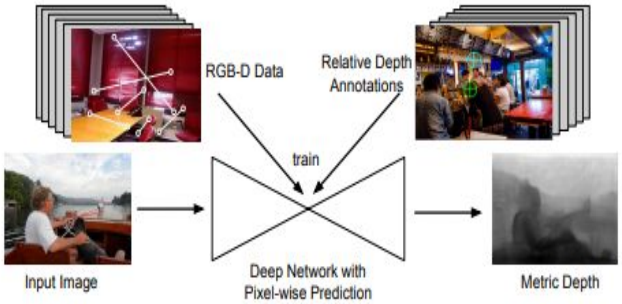
Retrieved from: https://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf , https://arxiv.org/pdf/1804.00607.pdf , https://arxiv.org/pdf/1606.00373.pdf

RW: Learning-based depth prediction

- Predicting dense depth from a single RGB image.
- Some methods also consider multiple images

None of them is designed to predict the depth of dynamic objects.

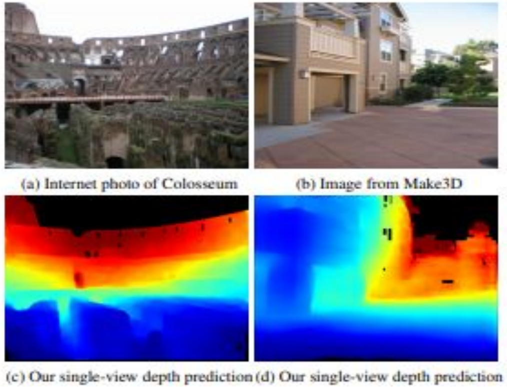
Single Image Depth Perception in the wild



RGB-D Data
Relative Depth Annotations
train
Input Image
Deep Network with Pixel-wise Prediction
Metric Depth

Figure 1: We crowdsource annotations of relative depth and train a deep network to recover depth from a single image taken in unconstrained settings ("in the wild").

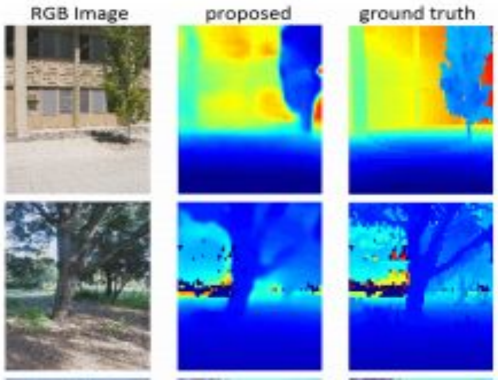
MegaDepth: Learning Single-View Depth Prediction from Internet Photos



(a) Internet photo of Colosseum (b) Image from Make3D

(c) Our single-view depth prediction (d) Our single-view depth prediction

Deeper Depth Prediction with Fully Convolutional Residual Networks



Retrieved from: https://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf , https://arxiv.org/pdf/1804.00607.pdf , https://arxiv.org/pdf/1606.00373.pdf

RW: **Depth estimation for dynamic scenes**

RGB data is used for 3D modeling, but only some estimate depth

- Reconstruct sparse geometry of a dynamic scene. (1&2)
- Predicts depth of of moving soccer players using synthetic training data from FIFA video games. (3)

1. Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes

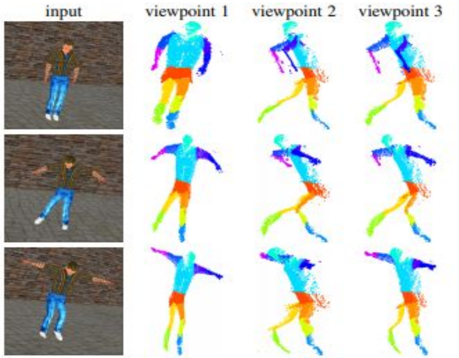
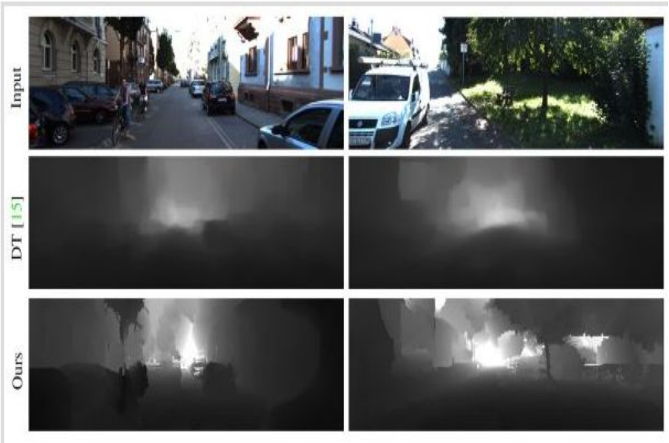


Fig. 10: Failure orthographic reconstruction results on dancer sequence. From left to right, we show input image and reconstruction results from three different viewpoints. Results are generated

2. Dense Monocular Depth Estimation in Complex Dynamic Scenes



3. Soccer on Your Tabletop

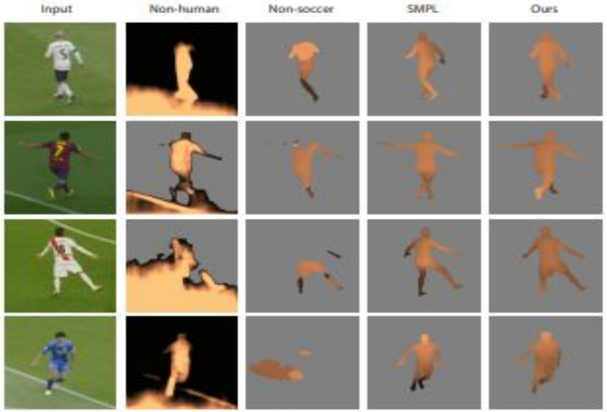


Figure 6. Results on real images from YouTube videos.

Retrieved From: http://www0.cs.ucl.ac.uk/staff/R.Yu/video_popup/VideoPopup_pami-compressed.pdf , http://vladlen.info/publications/dense-monocular-depth-estimation-in-complex-dynamic-scenes/ , http://www.krematas.com/Publications/soccer_on_your_tabletop.pdf

RW: Depth estimation for dynamic scenes

RGB data is used for 3D modeling, but only some estimate depth

- Reconstruct sparse geometry of a dynamic scene. (1&2)
- Predicts depth of of moving soccer players using synthetic training data from FIFA video games. (3)

1. **These methods impose strong assumptions of the object's motion which violates normal human motion.**
2. **Or these methods are very case specific.**

1. Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes

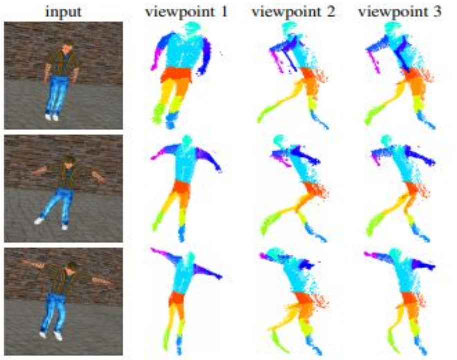
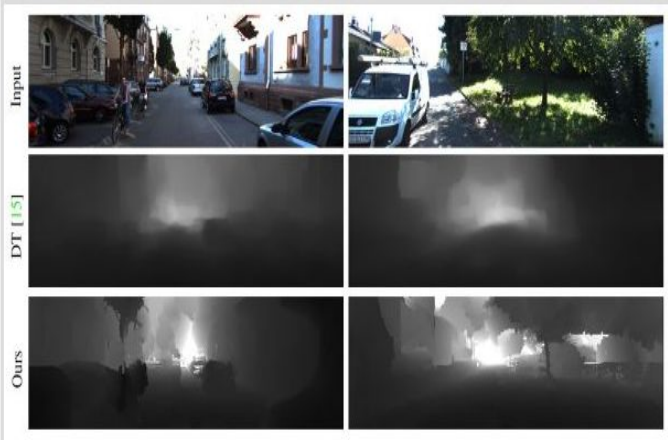


Fig. 10: Failure orthographic reconstruction results on dancer sequence. From left to right, we show input image and reconstruction results from three different viewpoints. Results are generated

2. Dense Monocular Depth Estimation in Complex Dynamic Scenes



3. Soccer on Your Tabletop

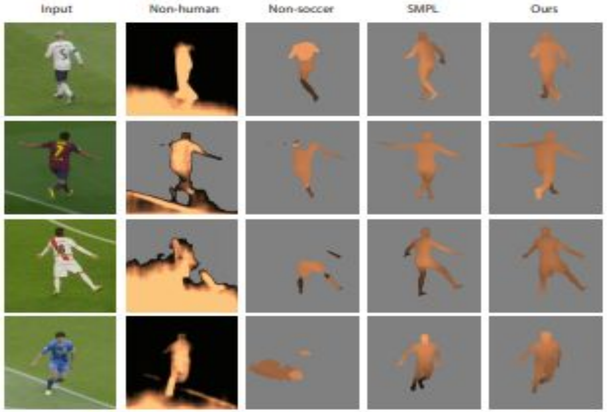


Figure 6. Results on real images from YouTube videos.

Retrieved From: http://www0.cs.ucl.ac.uk/staff/R.Yu/video_popup/VideoPopup_pami-compressed.pdf , http://vladlen.info/publications/dense-monocular-depth-estimation-in-complex-dynamic-scenes/ , http://www.krematas.com/Publications/soccer_on_your_tabletop.pdf

RW: **RGBD data for learning depth**

Capturing indoor scenes using depth sensors:

Indoor Segmentation and Support Inference from RGBD Images

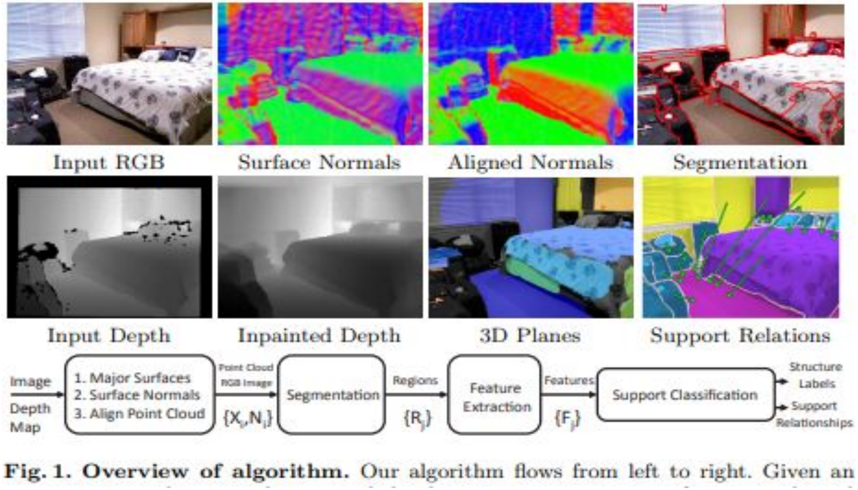


Fig. 1. Overview of algorithm. Our algorithm flows from left to right. Given an

REFRESH is a recent semi-synthetic scene flow dataset created by overlaying animated people:

Learning Rigidity in Dynamic Scenes for 3D Motion Field Estimation

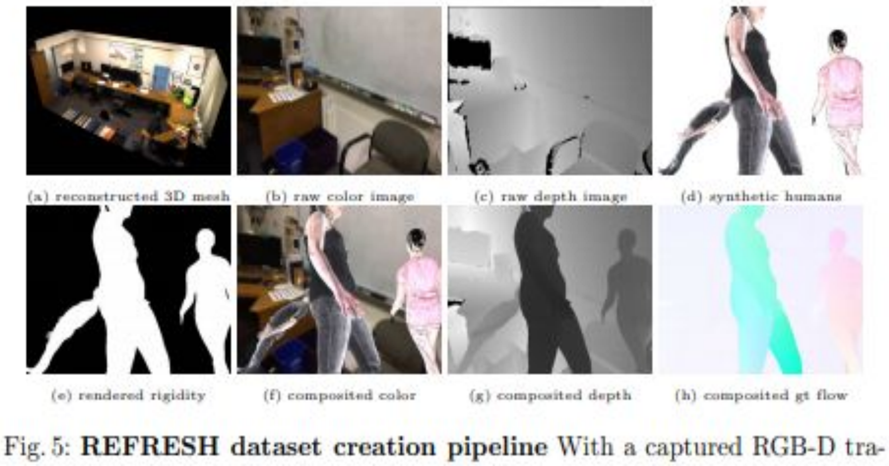


Fig. 5: **REFRESH dataset creation pipeline** With a captured RGB-D tra-

Retrieved From: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/shkf_eccv2012.pdf , https://arxiv.org/pdf/1804.04259.pdf

RW: **RGBD data for learning depth**

Dataset limited to indoor scenes and consists of synthetic humans placed in unrealistic configuration.

Capturing indoor scenes using depth sensors:

Indoor Segmentation and Support Inference from RGBD Images

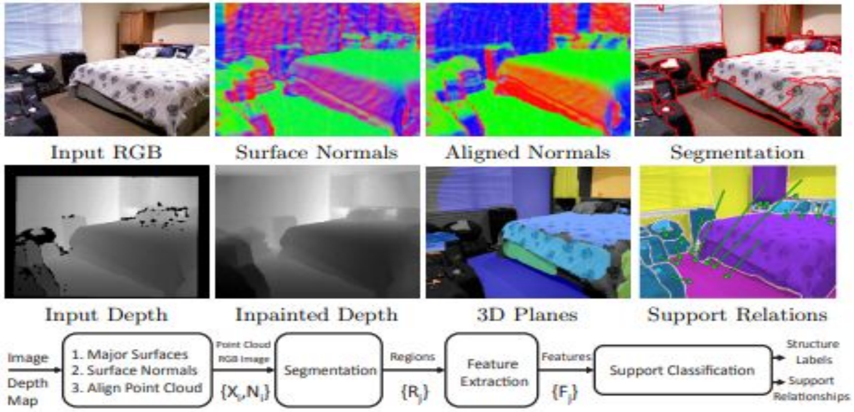


Fig. 1. Overview of algorithm. Our algorithm flows from left to right. Given an

REFRESH is a recent semi-synthetic scene flow dataset created by overlaying animated people:

Learning Rigidity in Dynamic Scenes for 3D Motion Field Estimation

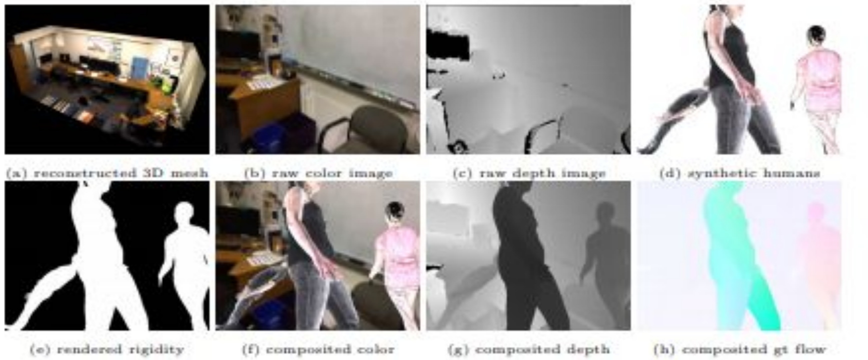


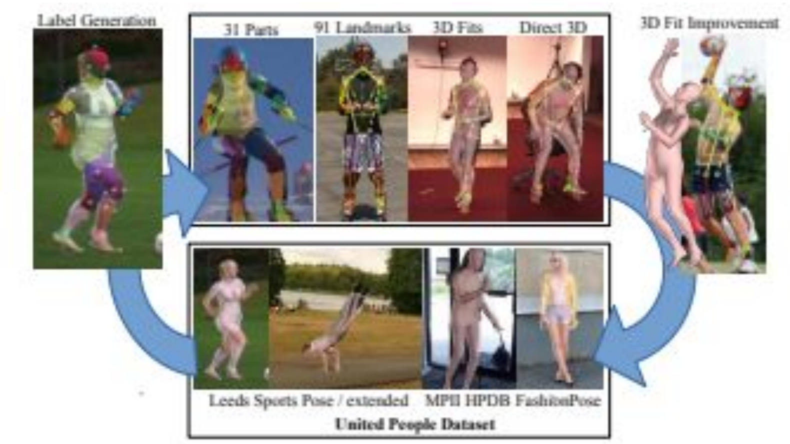
Fig. 5: **REFRESH dataset creation pipeline** With a captured RGB-D tra-

Retrieved From: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/shkf_eccv2012.pdf , https://arxiv.org/pdf/1804.04259.pdf

RW: Human shape and pose prediction

Natural image spanning on variety of poses by recovering posed 3D human mesh from a single RGB.

Unite the People: Closing the Loop Between 3D and 2D Human Representations



Automatic Estimation of 3D Human Pose and Shape from a Single Image

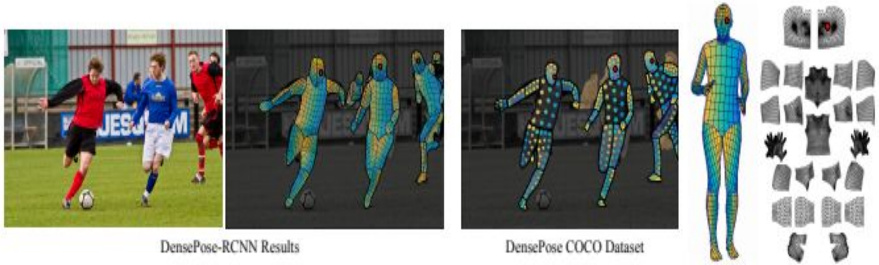


Figure 1: Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. We

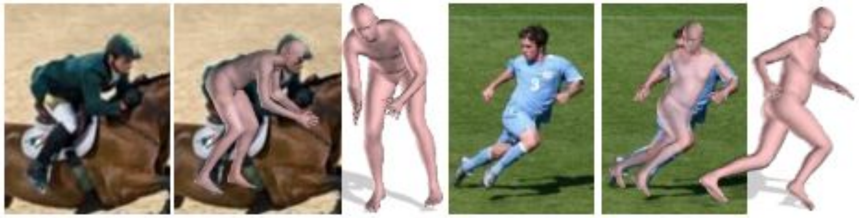


Fig. 1. Example results. 3D pose and shape estimated by our method for two images from the Leeds Sports Pose Dataset [22]. We show the original image (left), our fitted model (middle), and the 3D model rendered from a different viewpoint (right).

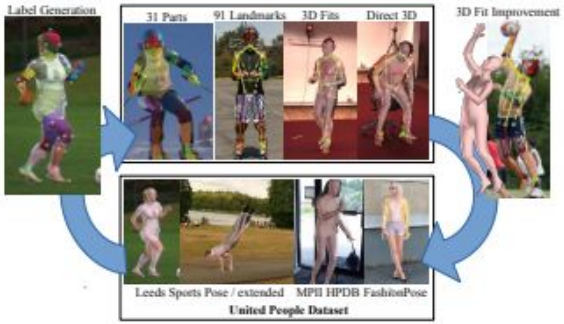
Retrieved From: https://arxiv.org/pdf/1607.08128.pdf , https://arxiv.org/pdf/1802.00434.pdf , https://arxiv.org/pdf/1701.02468.pdf

RW: Human shape and pose prediction

Natural image spanning on variety of poses by recovering posed 3D human mesh from a single RGB.

1. **Only model human body, disregarding hair, clothing, & non-human part of the scenes.**
2. **Rely on keypoints: require most of the body to be within frame**

Unite the People: Closing the Loop Between 3D and 2D Human Representations



Automatic Estimation of 3D Human Pose and Shape from a Single Image

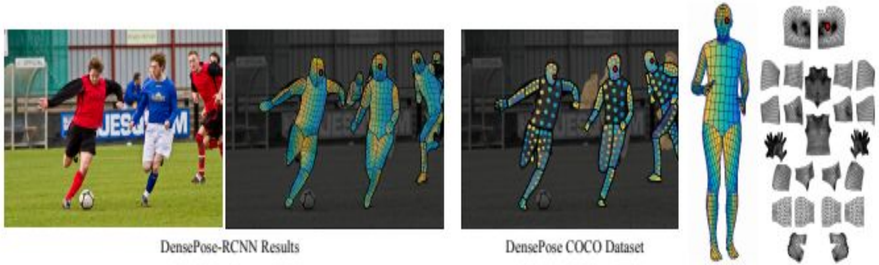


Figure 1: Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. We

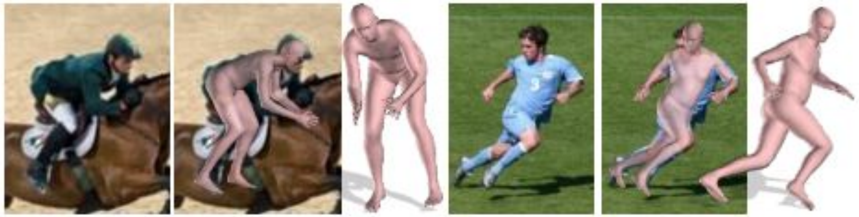


Fig. 1. Example results. 3D pose and shape estimated by our method for two images from the Leeds Sports Pose Dataset [22]. We show the original image (left), our fitted model (middle), and the 3D model rendered from a different viewpoint (right).

Retrieved From: https://arxiv.org/pdf/1607.08128.pdf , https://arxiv.org/pdf/1802.00434.pdf , https://arxiv.org/pdf/1701.02468.pdf

New Mannequin Challenge (MC) Dataset

Data is derived from:

Youtube videos in which people imitate mannequins, i.e. freeze in elaborate, natural poses, while a hand-held camera films the scene by moving around them.

New Mannequin Challenge (MC) Dataset

1. Thousands of such videos have been created and uploaded on Youtube since 2016, out of which about 2,000 are possible candidate videos for processing.
2. These videos span a wide range of scenes with people of different ages, naturally posing.

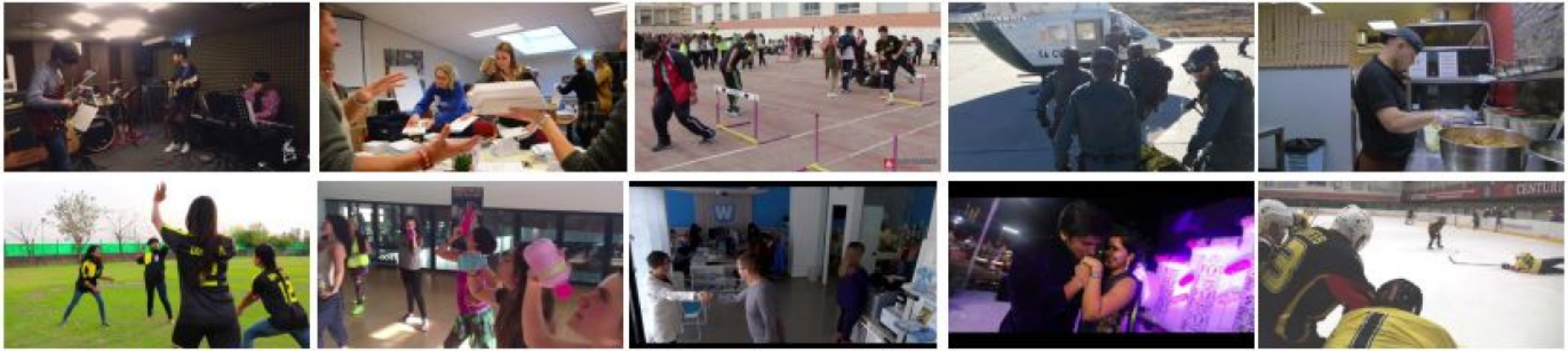


Figure 2. **Sample images from Mannequin Challenge videos**. Each image is a frame from a video sequence in which the camera is moving but *humans are all static*. The videos span a variety of natural scenes, poses, and configuration of people.

Estimating Camera Poses on MC Dataset

01

ORB-SLAM2

- Identify trackable sequences in each video
- Estimate an initial camera pose for each frame

- Process a low-resolution version of video for efficiency
- Set the field view to be 60 degrees

Estimating Camera Poses on MC Datset

01 **ORB-SLAM2**

- Identify trackable sequences in each video
- Estimate an initial camera pose for each frame

- Process a low-resolution version of video for efficiency
- Set the field view to be 60 degrees

02 **SfM Visual System**

- Extracts and matches features across frames
- Performs a global bundle adjustment optimization

- Reprocess video at a higher resolution.
- Refines the initial camera poses and intrinsic parameters

Estimating Camera Poses on MC Dataset

01 **ORB-SLAM2**

- Identify trackable sequences in each video
- Estimate an initial camera pose for each frame

- Process a low-resolution version of video for efficiency
- Set the field view to be 60 degrees

02 **SfM Visual System**

- Extracts and matches features across frames
- Performs a global bundle adjustment optimization

- Reprocess video at a higher resolution.
- Refines the initial camera poses and intrinsic parameters

03 **Non-smooth camera motion**

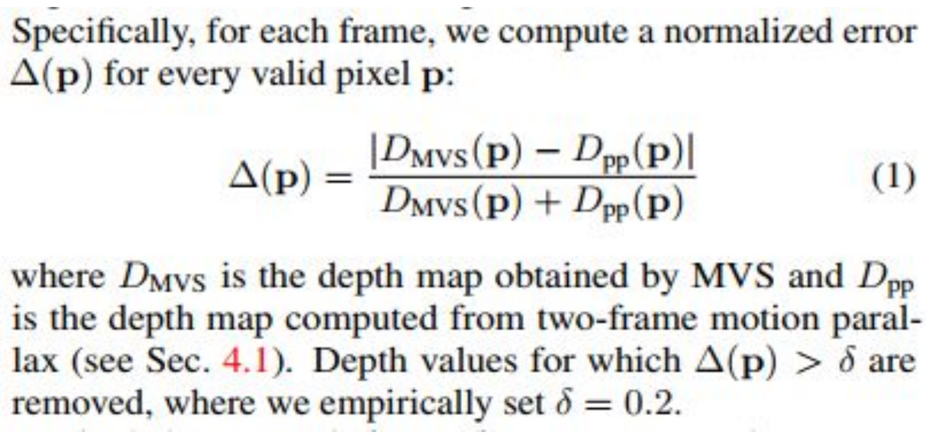
- Removing sequences with non-smooth camera motion

Depth Filtering Mechanism for MC Dataset

Raw depth maps generated by MVS result in excessive noise (like camera motion blur, shadows, etc.).

A careful depth filtering mechanism is adopted to deal with noise issue.

For each frame, a normalized error **(Δ p)** is computed for every **pixel p**.



Specifically, for each frame, we compute a normalized error $\Delta(\mathbf{p})$ for every valid pixel $\mathbf{p}$:

$$\Delta(\mathbf{p}) = \frac{|D_{\mathrm{MVS}}(\mathbf{p}) - D_{\mathrm{pp}}(\mathbf{p})|}{D_{\mathrm{MVS}}(\mathbf{p}) + D_{\mathrm{pp}}(\mathbf{p})} \tag{1}$$

where D_{MVS} is the depth map obtained by MVS and D_{pp} is the depth map computed from two-frame motion parallax (see Sec. 4.1). Depth values for which $\Delta(\mathbf{p}) > \delta$ are removed, where we empirically set $\delta = 0.2$.

Filtering Clips - Unsuitable Factors

MVS method is used to filter out the data that **do not obey** multi-view geometric constraints.

Several factors that makes video clip unsuitable for training:

1. People may “unfreeze” (**start moving**) at some point
2. Video may contain **synthetic graphical elements** in the **background**

Filtering Clips - Frames Filtered

Thus, these frames are filtered out by MVS:

1. Frames with **<20% of pixels** have valid MVS depth after two pass cleaning stage
2. Frames with estimated radical distortion coefficient **|k1| > 0.1 (fisheye camera)**
3. Frames with estimated focal length of <= 0.6 or >= 1.2

Filtering Clips - Sequences kept

Sequences kept:

1. At least **30 frames** long
2. Have an **aspect ratio** of **16:9**
3. Have a **width** of **>= 1600 pixels**

Final inspection is done **manually** removing obvious incorrect reconstructions.
A total of **4,690 sequences** are obtained with a total of more than 170K valid image-depth pairs.

The Depth Prediction Model

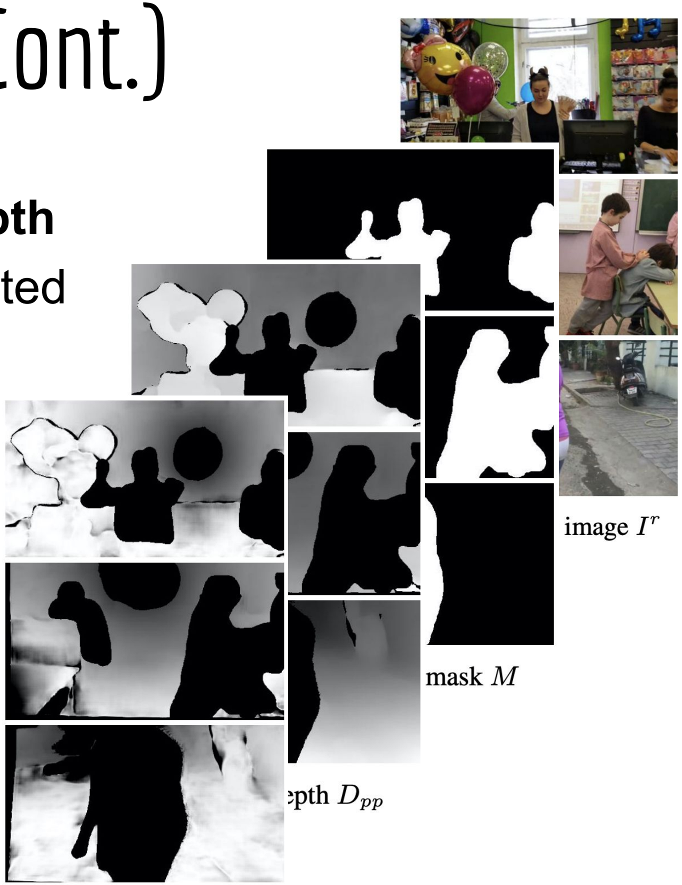
It’s a **supervised learning** model

The architecture derives from the **“hourglass” network** from “*Single-image depth perception in the wild.*” with the ***nearest-neighbor upsampling layers*** replaced by **bilinear upsampling layers**.

The Depth Prediction Model (Cont.)

It takes **four inputs and an MVS computed depth map** and is trained to regress to the MVS computed depth map

- A reference image denoted as $\boldsymbol{I}^r$
- A binary mask of human regions denoted as $\boldsymbol{M}$
- A depth map of the static background with human regions removed denoted as $\boldsymbol{D}_{pp}$
- A confidence map $\boldsymbol{C}$



(d) Input confidence C

Architecture

It could output the same resolution as the input, and at the upsampling stages it adds back features from high resolutions.

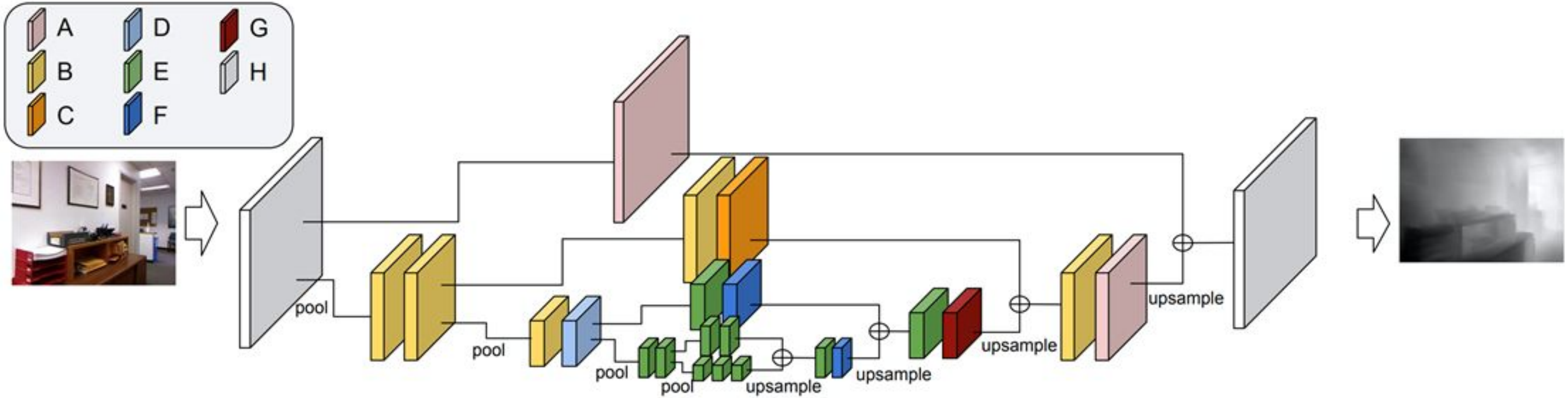
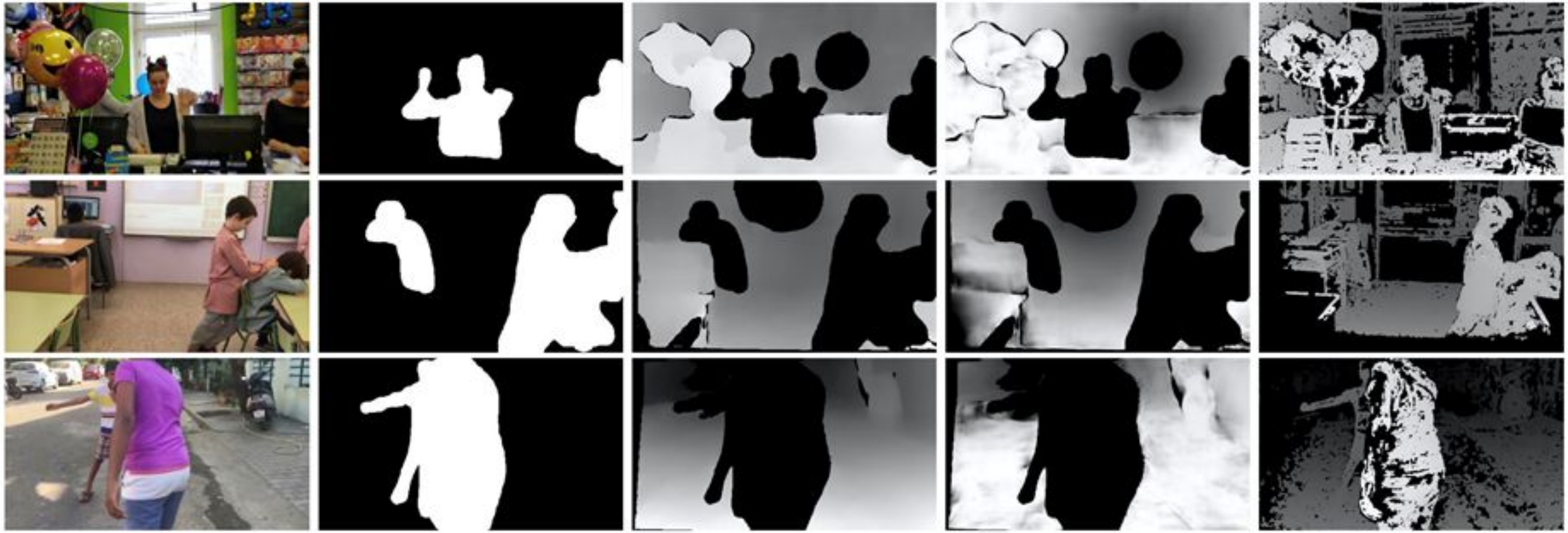


Image retrieved from: https://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf

Inputs Overview



(a) Reference image I^r (b) Human mask M (c) Input depth D_{pp} (d) Input confidence C MVS depth D_{MVS}

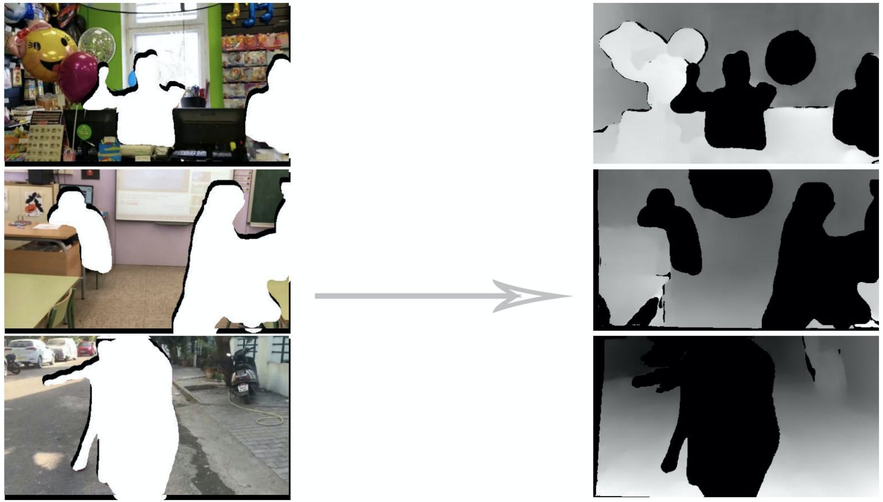
Ground Truth

The Initial Depth Map

Purpose:

to gather initial depth information of the background from the motion parallax between two frames so that could get accurate depth prediction for the background unavailable in a single frame

Assumption: Background is static while the humans are dynamic



(a) Reference image I^r

(c) Input depth D_{pp}

The Initial Depth Map (Cont.)

Method: Estimate an optical flow from reference image $\boldsymbol{I^r}$ and source image $\boldsymbol{I^s}$.

Frames Selection Method: Choose a source frame with significant overlap but large enough baseline with the reference frame

- **$s(index)=\arg\max d^{rj}\ o^{rj}$**

Confidence Map

Purpose: To counter the noises in the Internet video clips

Equation: $C(p)=C_{lr}(\mathbf{p})*C_{ep}(\mathbf{p})*C_{pa}(\mathbf{p})$

- $C_{lr}(\mathbf{p}) \; = \; \max\left(0, 1 - r(\mathbf{p})^2\right)$

  measures the left-right consistency between the forward and backward flow field.

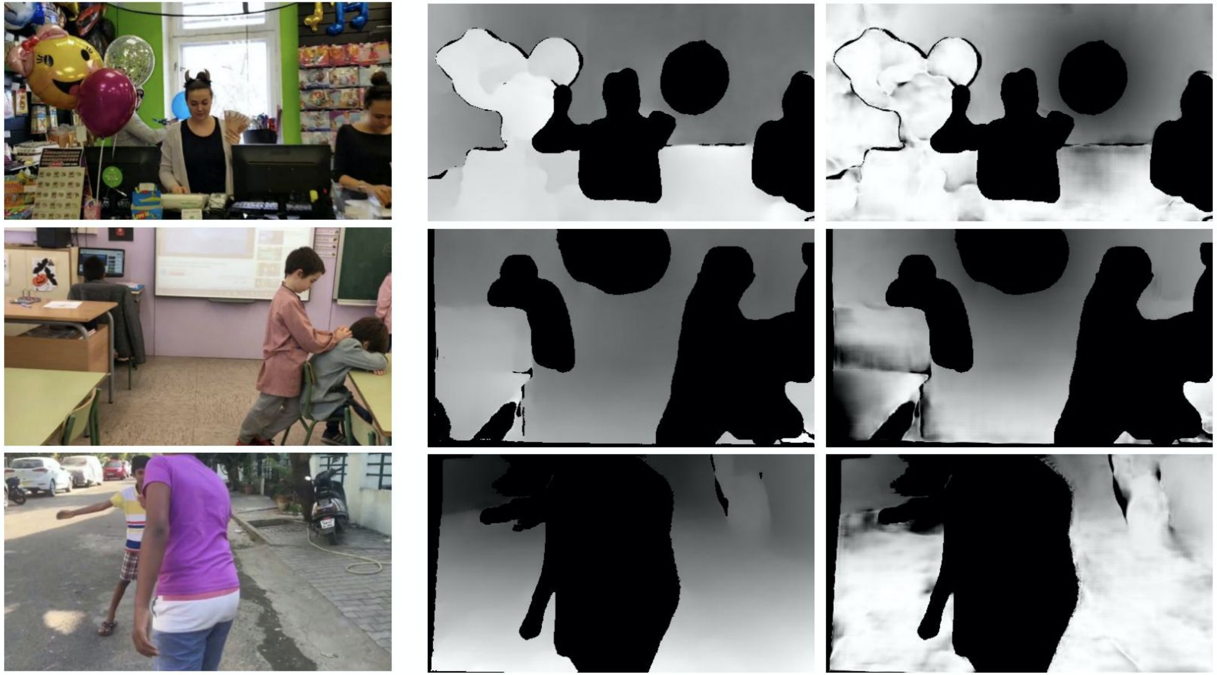
- $C_{ep}(\mathbf{p}) \; = \; \max\left(0, 1 - (\gamma(\mathbf{p})/\bar{\gamma})^2\right)$

  measures how well the flow field complies with the epipolar constraint between the views

- $C_{pa}(\mathbf{p}) = 1 - \left(\frac{\min(\bar{\beta}, \beta(\mathbf{p})) - \bar{\beta}}{\bar{\beta}}\right)^2$

  Assigns low confidence to pixels for which the parallax between the views is small

Confidence Map (Cont.)



(a) Reference image I^r (c) Input depth D_{pp} (d) Input confidence C

Loss

Key features: scale-invariant

$L_{si}=L_{MSE}+a1*L_{grad}+a2*L_{sm}$

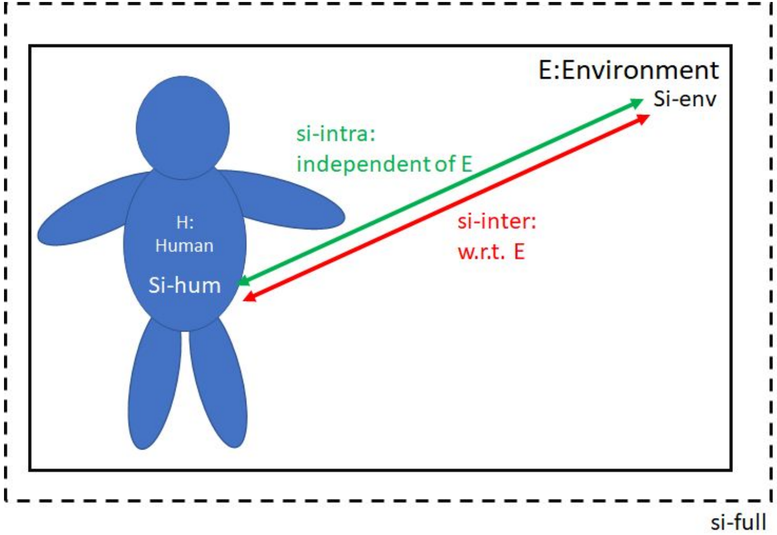
- L_{MSE} is the scale-invariant mean square error.
- L_{grad} is to recover the sharp depth discontinuities and make the gradient change matched to the ground truth.
- L_{sm} is a smooth interpolation of depth in texture-less regions where MVS fails to recover depth in the ground truth

Error metrics

We measure error using the scale-invariant RMSE (si-RMSE), equivalent to the square of L_{MSE}.

We evaluate si-RMSE on 5 different regions.

Lower error metric, better performance!



Evaluation on MC test set

- If input an optical flow field to the network instead of depth, the performance is only comparable to the single view method.

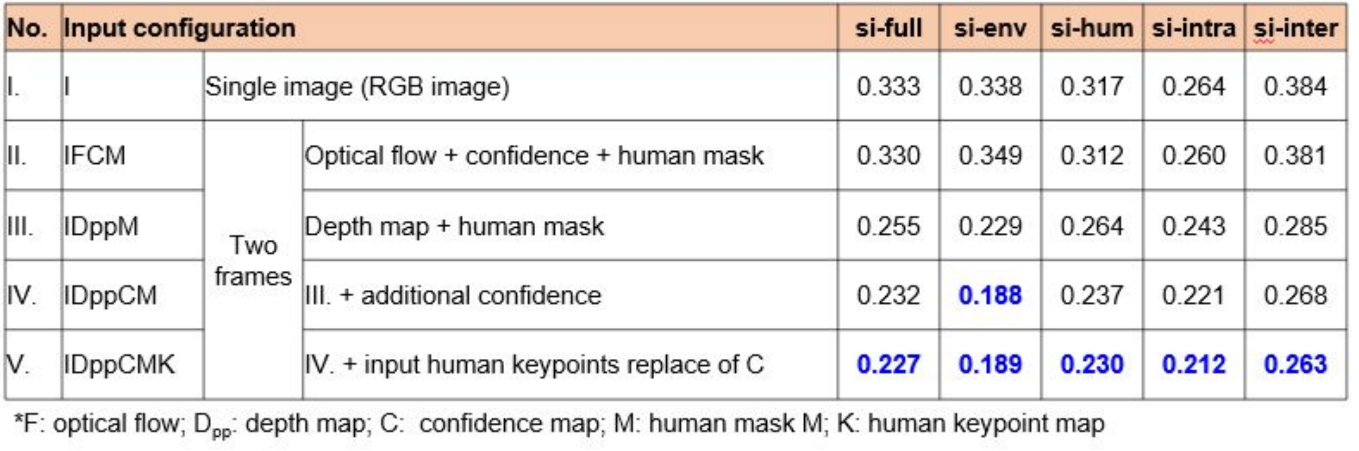
- I. & II.

- Adding the initial depth of environment and a confidence map can improve the performance for both human and non-human regions.

- I.,III. & IV.

- Adding human keypoint locations to the network input can further improve the performance.

- IV. & V.

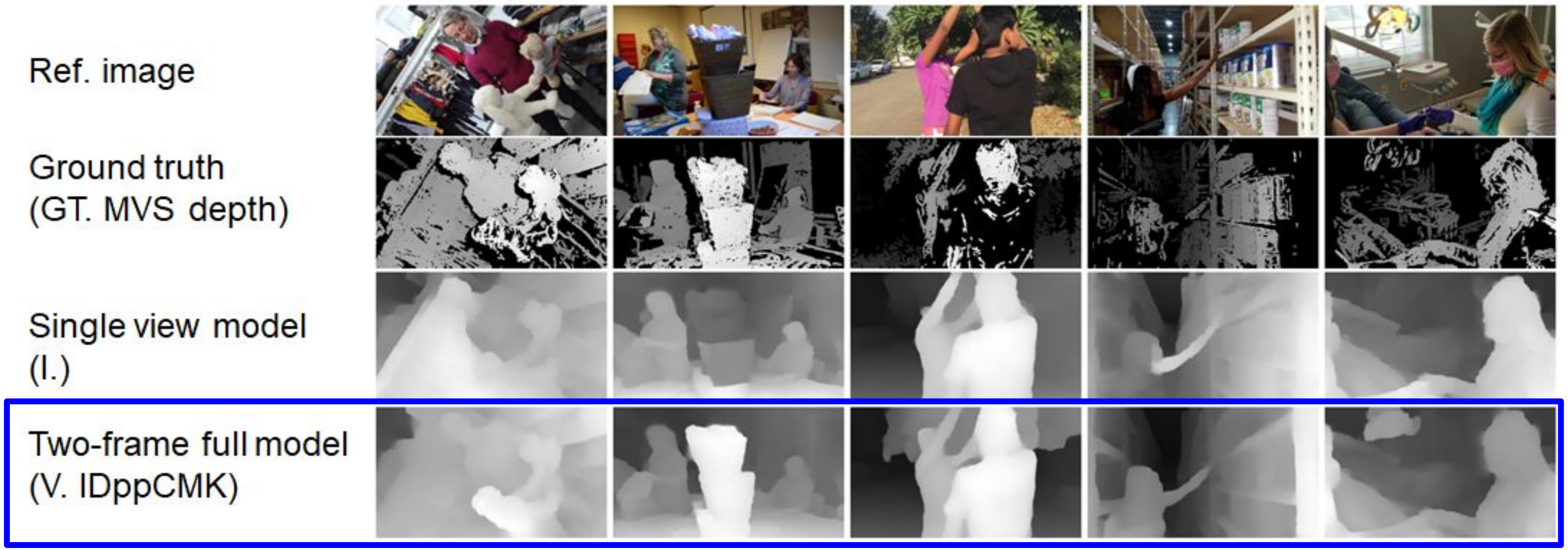


No.	Input configuration			si-full	si-env	si-hum	si-intra	si-inter
I.	I	Single image (RGB image)		0.333	0.338	0.317	0.264	0.384
II.	IFCM	Two frames	Optical flow + confidence + human mask	0.330	0.349	0.312	0.260	0.381
III.	IDppM		Depth map + human mask	0.255	0.229	0.264	0.243	0.285
IV.	IDppCM		III. + additional confidence	0.232	**0.188**	0.237	0.221	0.268
V.	IDppCMK		IV. + input human keypoints replace of C	**0.227**	**0.189**	**0.230**	**0.212**	**0.263**

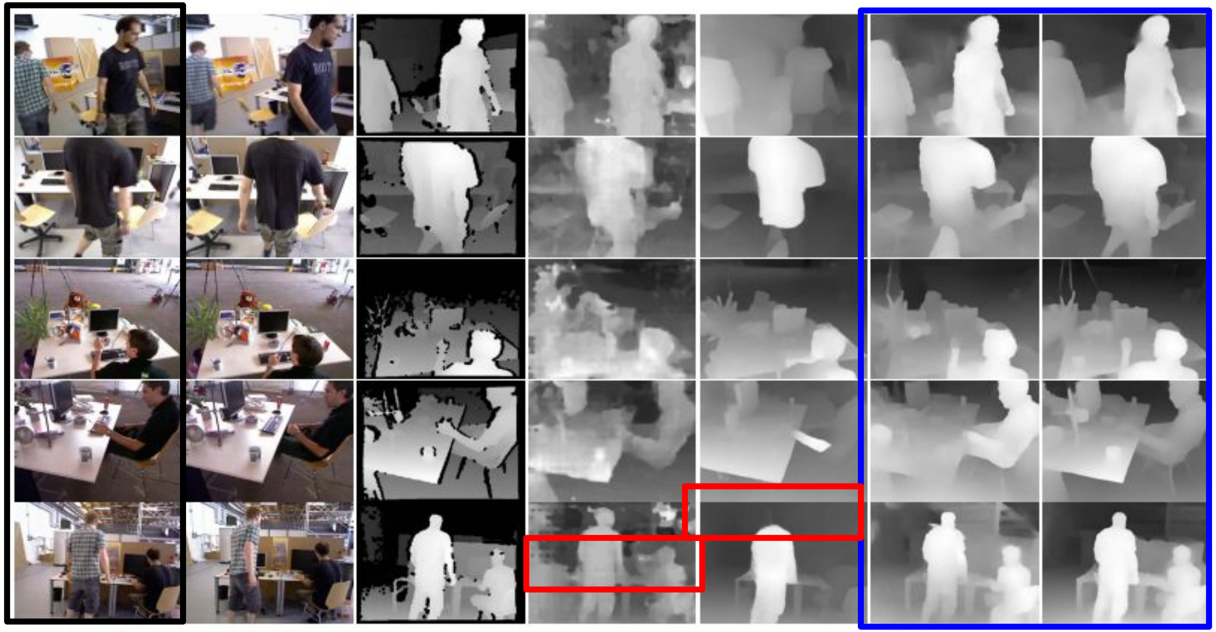
*F: optical flow; D_{pp}: depth map; C: confidence map; M: human mask M; K: human keypoint map

Evaluation on MC test set (Cont.)

Compared to single view model, the full model results (V. IDppCMK) are more accurate in both human regions and non-human regions .



Evaluation of TUM RGBD dataset



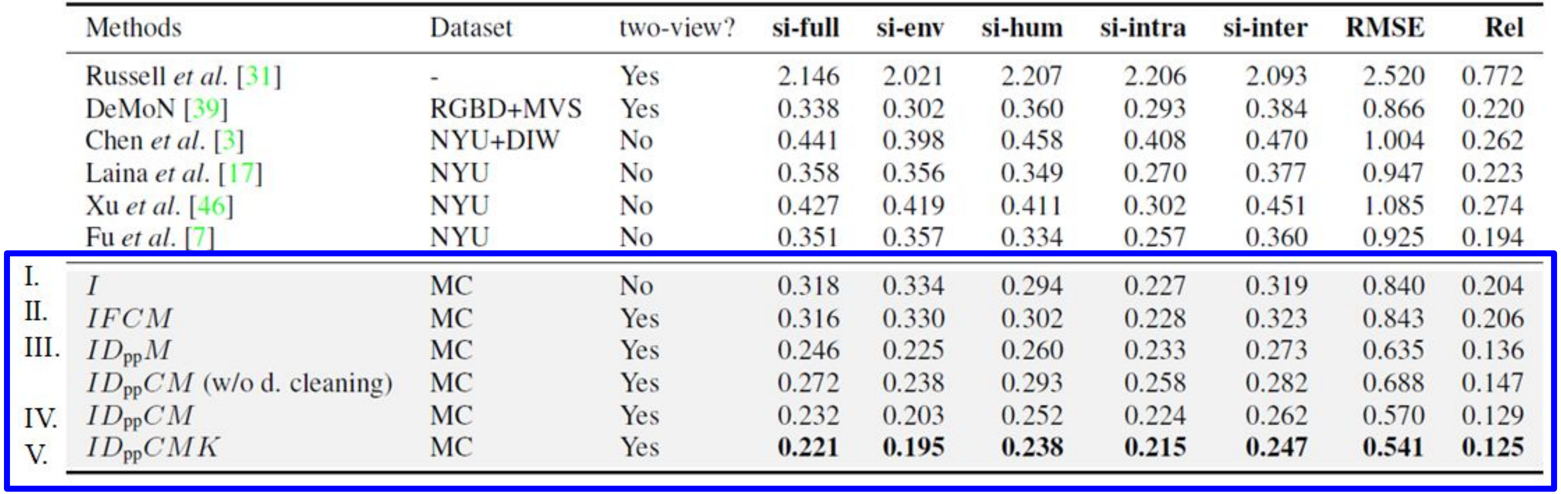
(a) I^r (b) I^s (c) GT (d) DORN [7] (e) DeMoN [39] (f) Ours (RGB) (g) Ours (full)

Figure 5. **Qualitative comparisons on TUM RGBD dataset.** (a) Reference images, (b) source images (used to compute our initial depth input), (c) ground truth sensor depth, (d) single view depth prediction method DORN [7], (e) two-frame motion stereo DeMoN [39], (f-g) depth predictions from our single view and two-frame models, respectively.

Our model's depth predictions strongly resemble the ground truth and show high level of details and sharpness.

Results on the TUM RGBD Dataset

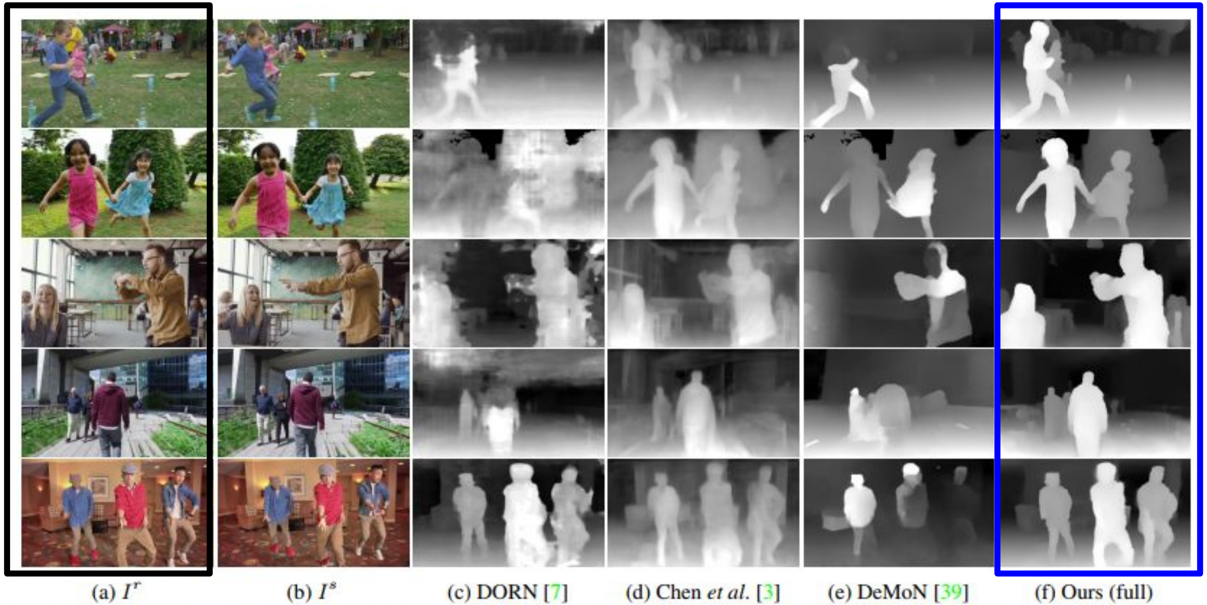
1. Our single-view model outperforms other single-view models due to training on MC dataset.
2. The full model significantly improves performance for all error measures.



	Methods	Dataset	two-view?	**si-full**	**si-env**	**si-hum**	**si-intra**	**si-inter**	**RMSE**	**Rel**
	Russell *et al.* [31]	-	Yes	2.146	2.021	2.207	2.206	2.093	2.520	0.772
	DeMoN [39]	RGBD+MVS	Yes	0.338	0.302	0.360	0.293	0.384	0.866	0.220
	Chen *et al.* [3]	NYU+DIW	No	0.441	0.398	0.458	0.408	0.470	1.004	0.262
	Laina *et al.* [17]	NYU	No	0.358	0.356	0.349	0.270	0.377	0.947	0.223
	Xu *et al.* [46]	NYU	No	0.427	0.419	0.411	0.302	0.451	1.085	0.274
	Fu *et al.* [7]	NYU	No	0.351	0.357	0.334	0.257	0.360	0.925	0.194
I.	I	MC	No	0.318	0.334	0.294	0.227	0.319	0.840	0.204
II.	$IFCM$	MC	Yes	0.316	0.330	0.302	0.228	0.323	0.843	0.206
III.	$ID_{\text{pp}}M$	MC	Yes	0.246	0.225	0.260	0.233	0.273	0.635	0.136
	$ID_{\text{pp}}CM$ (w/o d. cleaning)	MC	Yes	0.272	0.238	0.293	0.258	0.282	0.688	0.147
IV.	$ID_{\text{pp}}CM$	MC	Yes	0.232	0.203	0.252	0.224	0.262	0.570	0.129
V.	$ID_{\text{pp}}CMK$	MC	Yes	**0.221**	**0.195**	**0.238**	**0.215**	**0.247**	**0.541**	**0.125**

Evaluation on Internet videos of dynamic accesses

Our predicted depth maps demonstrate accurate depth ordering of both between the people and other objects in the scene.

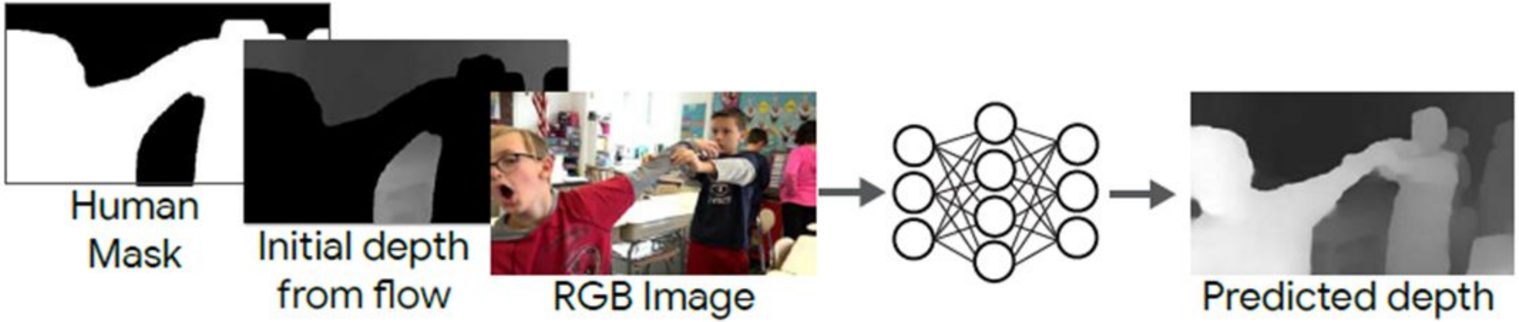


(a) I^r (b) I^s (c) DORN [7] (d) Chen *et al.* [3] (e) DeMoN [39] (f) Ours (full)

Figure 6. **Comparisons on Internet video clips with moving cameras and people.** From left to right: (a) reference image, (b) source image, (c) DORN [7], (d) Chen *et al.* [3], (e) DeMoN [39], (f) our full method.

Summary - Contributions & Technical Ideas

1. Using a new data source called "Mannequin Challenge Dataset"
2. A deep-network-based model is designed and trained to predict dense depth maps
   - Especially for the challenging case: simultaneous camera motions and complex human motions.

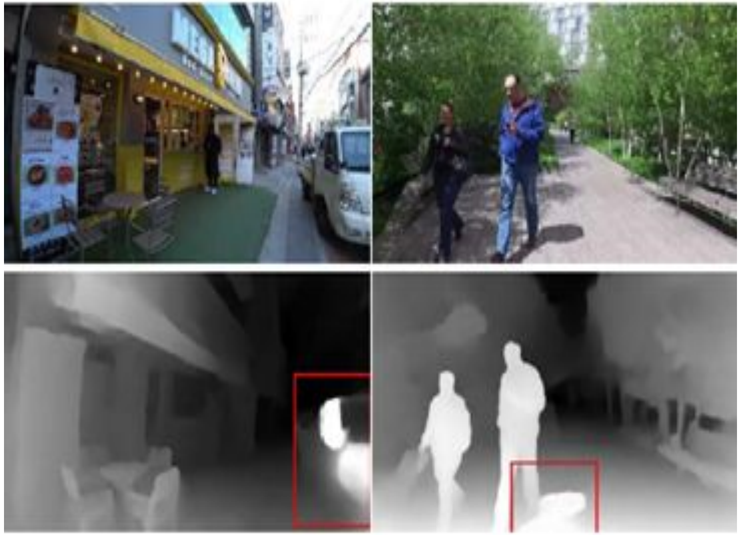


Summary - Strengths

Obtain a reliable depth supervision from such noisy data, and significantly outperform state-of-the-art methods.

Summary - Weaknesses

- Difficult to infer if moving objects cover most of the scene.
- The predicted depth may be inaccurate for non-human, moving regions such as cars and shadows
- Only use two views, sometimes leading to temporally inconsistent depth estimation.



Work Extension

A range of depth-visual effects such as

1. Depth-based defocus
2. Insertion of synthetic 3D graphics
3. Removal of humans by inpainting

Thank you!