Shape-Assisted Multimodal Person Re-Identification

Ph.D. Dissertation Defense
Haidong Zhu

Advisor: Prof. Ram Nevatia (Chair)
Committee: Prof. Ulrich Neumann
Prof. Antonio Ortega

Apr. 25, 2024

Problem Definition

- ❑ Person Re-identification (Re-ID)
  - ❑ Identify the person based on their biometric information
  - ❑ Match the person in the probe (query) with examples in the template gallery

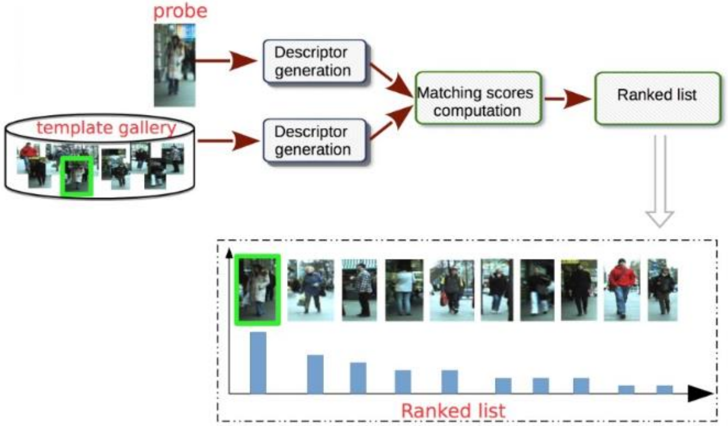


Image source: Lavi, Bahram, et al. "Survey on reliable deep learning-based person re-identification models: Are we there yet?." *arXiv preprint arXiv:2005.00355* (2020).

Challenges

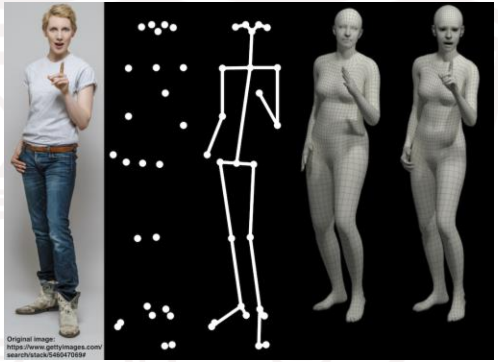
- ❑ Representation – single-frame *v.s.* video

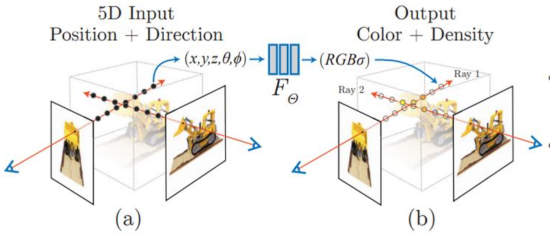
- ❑ Clothes conditions – same clothes *v.s.* clothes changes

- ❑ Quality – Occlusion, degradation, yaw/pitch angle, etc.

Motivation and Background

- 3-D representation, compared with 2-D images, has the strength of:
  - 3-D body shape invariant to variations
  - Include external body shape prior
  - Has significant development recently.





3-D shape Representation And Reconstruction

Curriculum DeepSDF (ECCV 2020)

Semantic Analysis for Training DeepSDF

General 3-D Shape Representation

CAT-NeRF (CVPRw 2023)

Shape Consistency across frames in a video

Animatable NeRF

Multimodal NeRF (ICRA 2023)

Point-cloud as Density Guidance for Training

Mulitmodal Analysis

Re-Identification

GaitHBS (WACV 2023)

Gait Recognition with 3-D Body Shape Guidance

Re-Identification → Gait with 3-D Shape

GaitRef (IJCB 2023)

Consistency between Skeletons and Shape

Re-Identification → Gait with 2-D Shape

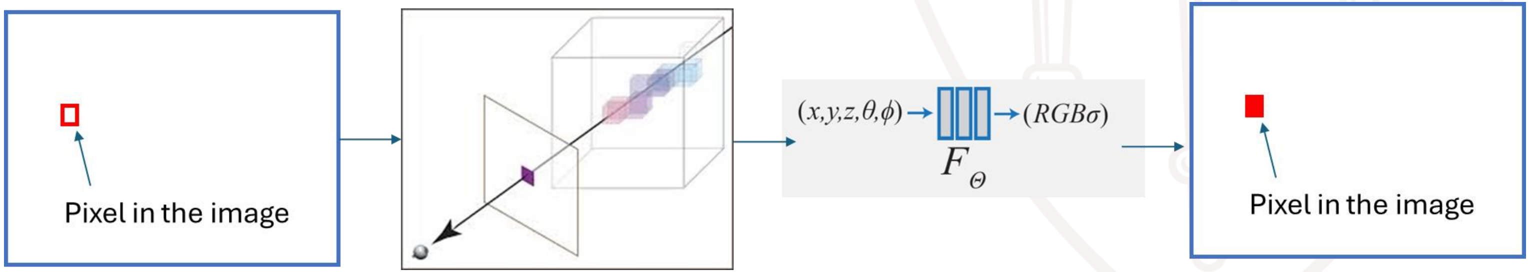
GaitSTR (TBIOM 2024)

Fusion between Shape and Different Skeletons

Re-Identification → Gait with 2-D Shape

3-D Reconstruction and Person Re-Identification

- NeRF-related representation mostly focuses on:
    - **Rendering quality** : Don’t mind if need to retrain for a new scene;
    - **Pixel-level accuracy**: Focus on pixel-level aggregation and rendering.



Using 3-D Shape Representation For Person Re-Identification

ShARc
(WACV 2024)

Multimodal Analysis for Person Re-Identification

Re-Identification → Multimodal Analysis

Comparing the Influence of Different Modalities for Whole-body Person Re-Identification

SEAS
(CVPR 2024)

Using 3-D Body Shape as Training Supervision

Re-Identification → Shape and Appearance

Using Body Shape as Shape-Aligned Guidance for Training Instead of Input

CaesarNeRF
(under review)

Semantic Extraction with NeRF using limited

3-D Representation → Semantic Analysis

Extracting 3-D Scene-level Representation with Generalizable NeRF Using Limited Reference Input

ShARc: Shape and Appearance Recognition For Person Identification In-the-wild

WACV 2024

Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, Ram Nevatia

Recognizing Person In-the-wild

- ❑ Videos captured in-the-wild suffers from:
  - ❑ Different activity between videos
  - ❑ Clothes variations
  - ❑ Atmosphere turbulence and degradations

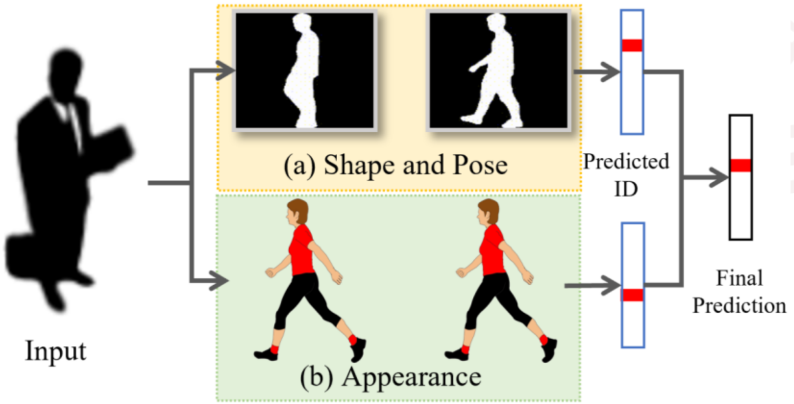
- ❑ Single whole-body modality cannot handle all variations

Gallery Frame	Standing Videos	Different Clothing	Turbulence & Occlusion
Gait		✓	✗
Body shape	✗	✗	✓
Appearance	✓	✗	✗

✓ Can be used for matching

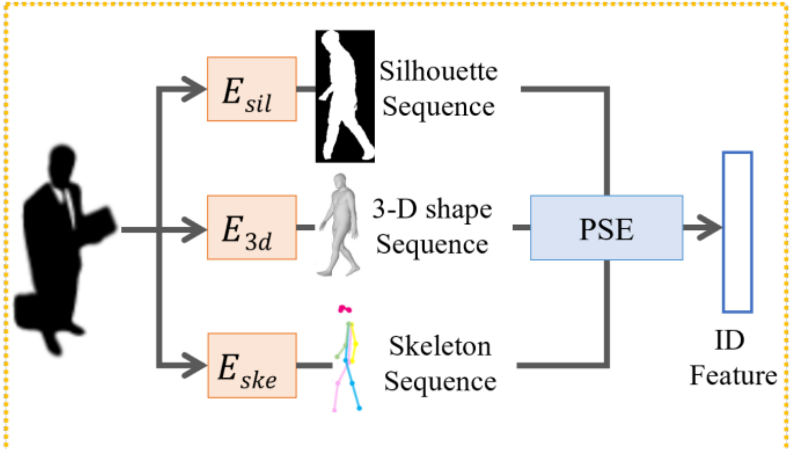
✗ Can be used for matching, but not very accurate

ShARc: Shape and Appearance Recognition

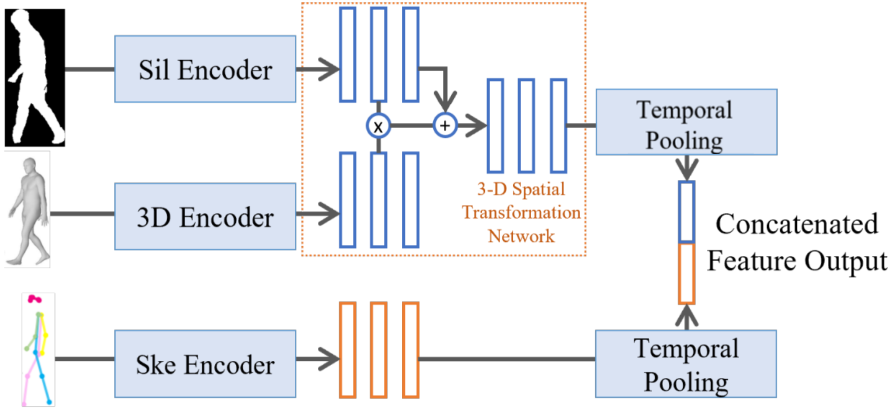


- ShARc decomposes the task to two branch with multimodality.
- Shape and pose recognize the person based on activity and body shape.
- Appearance focuses on directly fusing appearance across different frames.

PSE: Pose and Shape Encoder



(a) Pose and Shape Encoder (PSE)



PSE details

Extract silhouettes (2-D body shape), SMPL (3-D body shape) and skeletons (Pose) combine them correspondingly

PSE: Pose and Shape Encoder

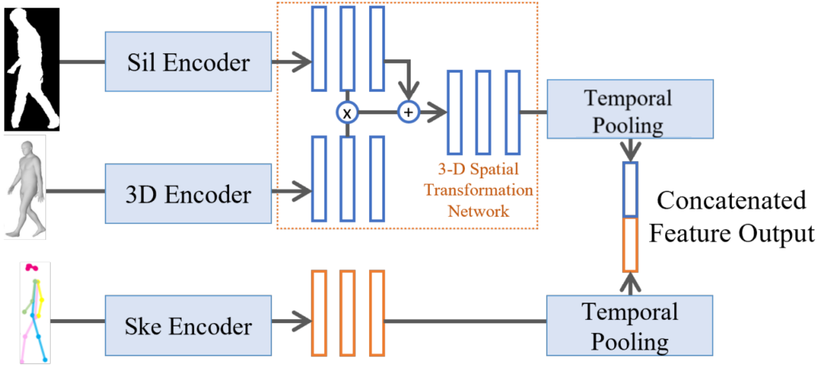
- 3-D Spatial Transformation Network

$$G_i = MLP(SMPL_i)$$

$$F_i = F_i \cdot (I + G_i)$$

$$F: 2D\ silhouette, G: 3D\ body\ feature$$

- Skeleton Encoder – STGCN



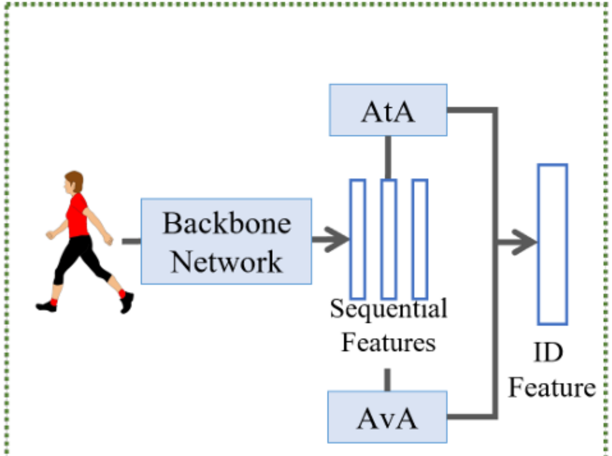
$$f_{out}(v_j) = \sum_{v_i \in B} \frac{1}{Z_{v_i}} f_{in}(v_i) \cdot w(v_i)$$

$$B = \{v_i | d(v_j, v_i) \leq K\ and\ \Delta t_{ij} < \tau\}$$

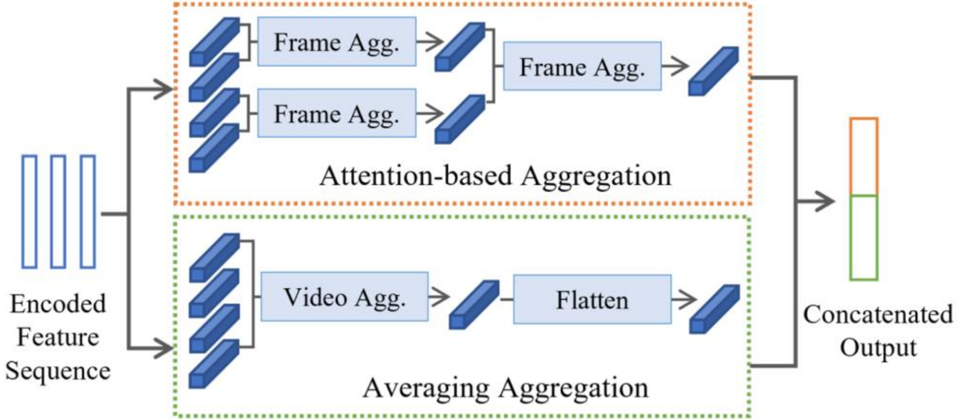
f_{in}: $feature\ input$ v: $vertex$ Z_v: $normalization\ factor$
τ: $temporal\ range\ hyperparam$ d: $distance\ of\ nodes$

Yan et al. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. AAAI 2018

AAE: Aggregated Appearance Encoder



(b) Agg. Appearance Encoder (AAE)



Two different ways of aggregating feature maps across different frames in the video for appearance recognition.

AAE: Aggregated Appearance Encoder

- Attention-based aggregation (AtA)

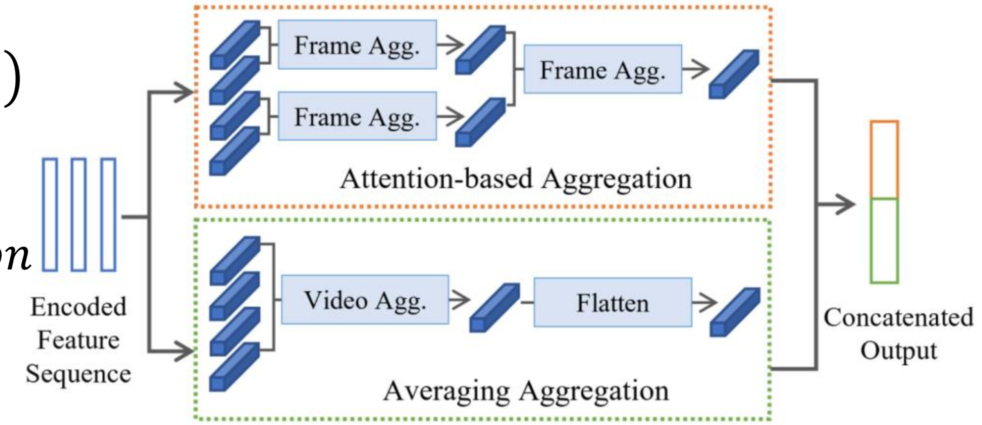
$$F_{t'}^{l+1} = w_1 \cdot F_t^l + w_2 \cdot F_{t+1}^l + w_3 \cdot S\left(F_t^l, F_{t+1}^l\right)$$

$$w_1 + w_2 + w_3 = 1$$

$F: feat\ map, l: level, t: timestamp, S: fusion$

- Averaging Aggregation (AvA)

$$F_{out} = \frac{1}{N} \sum F_t$$



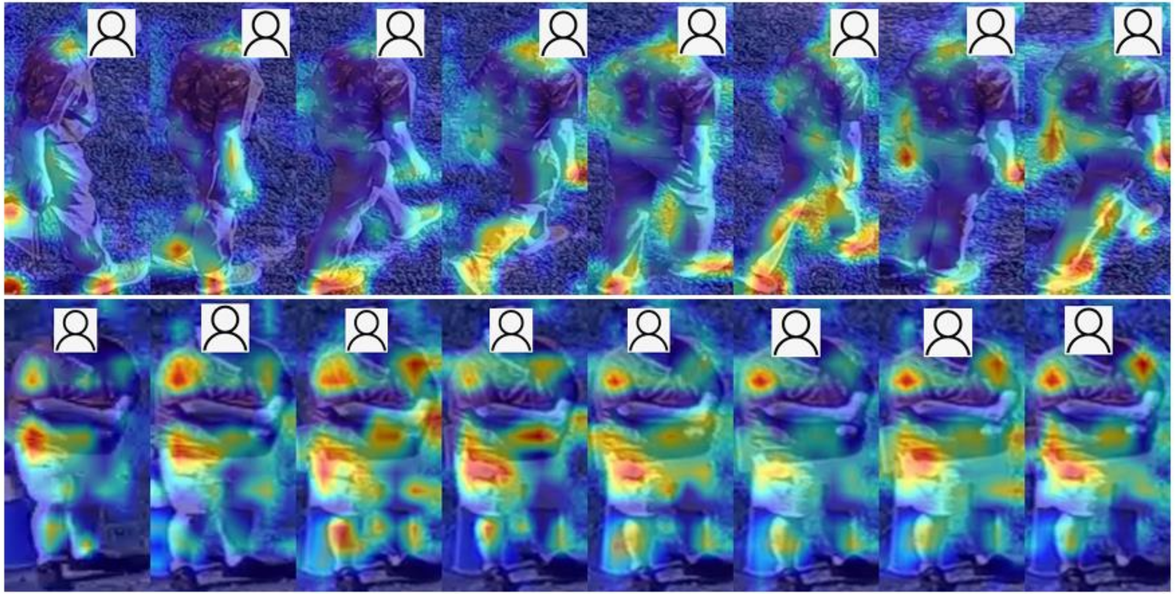
Main Result

Method	All Activities	
	Rank 1	Rank 20
GaitSet [2]	15.3	40.5
GaitPart [12]	14.1	41.7
GaitGL [35]	15.6	45.1
GaitMix [72]	15.9	46.5
GaitRef [72]	17.7	50.2
SMPLGait [66]	18.8	51.9
PSE (Ours)	21.2	65.3
DME [18]	25.0	63.8
PSTA [58]	33.6	67.3
CAL [16]	34.9	71.4
TCL Net [23]	31.3	65.6
Attn-CL+rerank [44]	27.6	61.8
AAE (Ours)	38.3	81.8
ShARc	**41.1**	**83.0**

PSE and AAE show state-of-the-art results for gait recognition and appearance recognition, while the aggregation, ShARc is the best

Cornett et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. WACV 2023

Main Result



When person is walking, attention map focuses on legs and arms, while it focuses more on shoulder or body area for stationary videos

USC

Cornett et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. WACV 2023

Main Result

Method	Rank-1	Rank-20
(Gait only) GaitRef [1]	17.7	50.2
(Gait + body) GaitHBS [2]	19.7	63.4
(App.) AAE [2]	38.3	81.8
(Gait + body + app.) ShARc [3]	41.1	83.0

- Body shape contributes limited improvement in the final pipeline
  - Body shapes are not directly combined with appearance
  - Body shapes are not accurate enough for recognizing the person

[1] Zhu* et al. Gaitref: Gait recognition with refined sequential skeletons. IJCB 2023
[2] Zhu et al. Gait recognition using 3-d human body shape inference. WACV 2023
[3] Zhu et al. Sharc: Shape and appearance recognition for person identification in-the-wild. WACV 2024.

SEAS: ShapE-Aligned Supervision for Person Reidentification

CVPR 2024

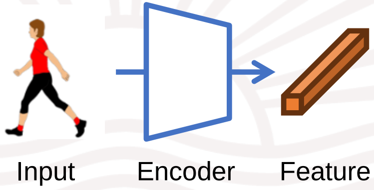
Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, Ram Nevatia

University of Southern California

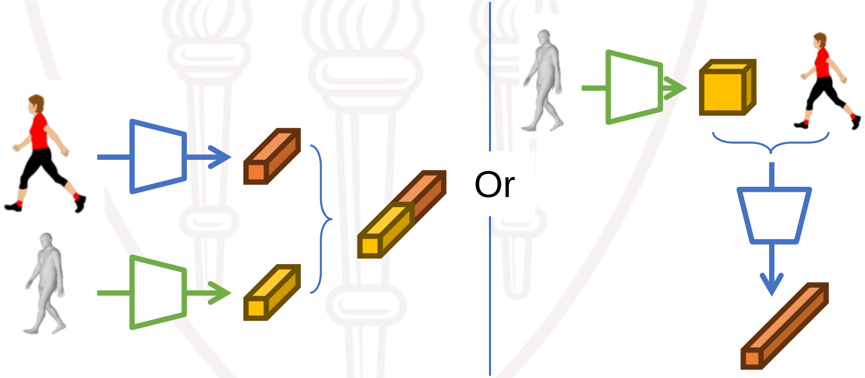
Existing Methods for Using 3-D Body shape

- Appearance-based recognition
  - Encode ID feature with an encoder;

- Using body shape as input
  - As a second branch [1];
  - As feature map [2] concatenated with RGB frames.



Appearance-based recognition

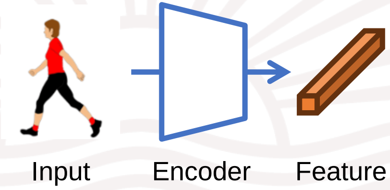


Shape and appearance recognition
(Two approaches)

USC

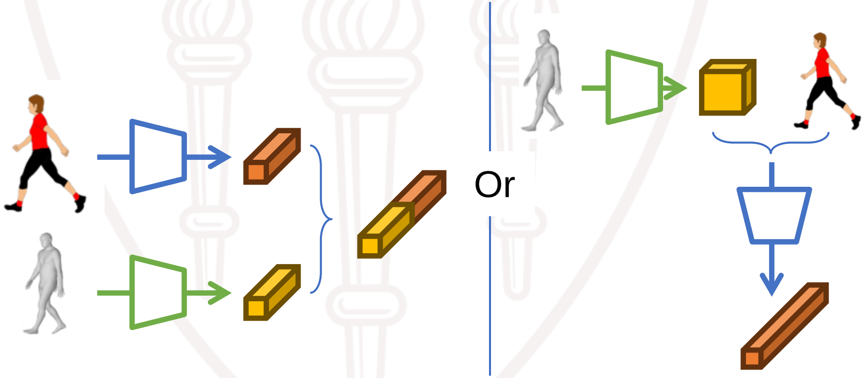
[1] Chen, Jiaxing, et al. "Learning 3D shape feature for texture-insensitive person re-identification." *CVPR*. 2021.
[2] Liu, Feng, et al. "Learning clothing and pose invariant 3d shape representation for long-term person re-identification." *ICCV*. 2023.

Existing Methods for Using 3-D Body shape



Appearance-based recognition

	Rank-1	mAP
Baseline	94.1	83.2
+ Shape as 2nd branch	94.1	**84.8**
+ Concatenated shape	**94.3**	**85.5**

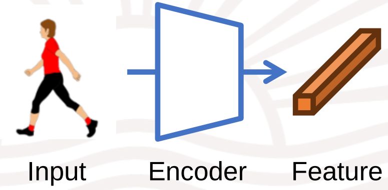


Shape and appearance recognition
(Two approaches)

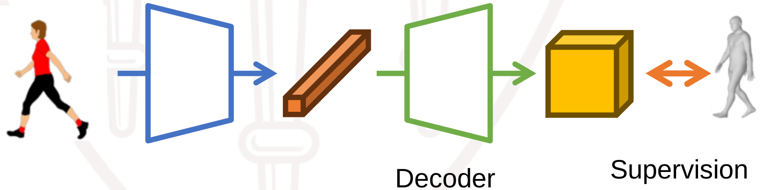
[1] Chen, Jiaxing, et al. "Learning 3D shape feature for texture-insensitive person re-identification." *CVPR*. 2021.
[2] Liu, Feng, et al. "Learning clothing and pose invariant 3d shape representation for long-term person re-identification." *ICCV*. 2023.

Using 3-D Body Shape as Supervision

- Using 3-D body shape as supervision ensure the body shape information is preserved in the encoded feature
- During inference, the decoder can be dropped, ensuring no extra computation cost for inference pipeline

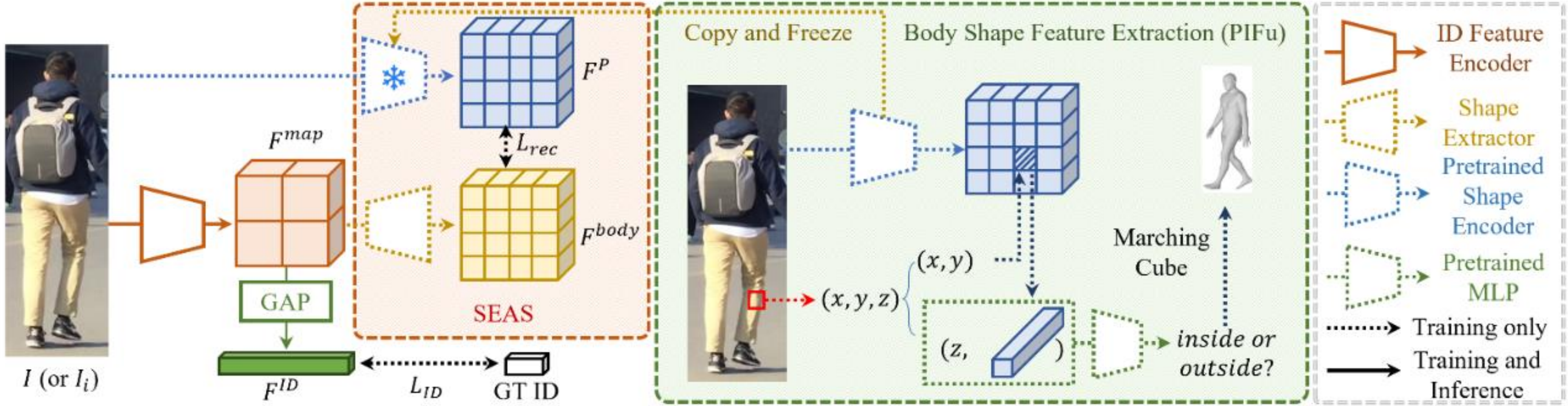


Appearance-based recognition



Shape As Supervision

Using 3-D Body Shape as Supervision



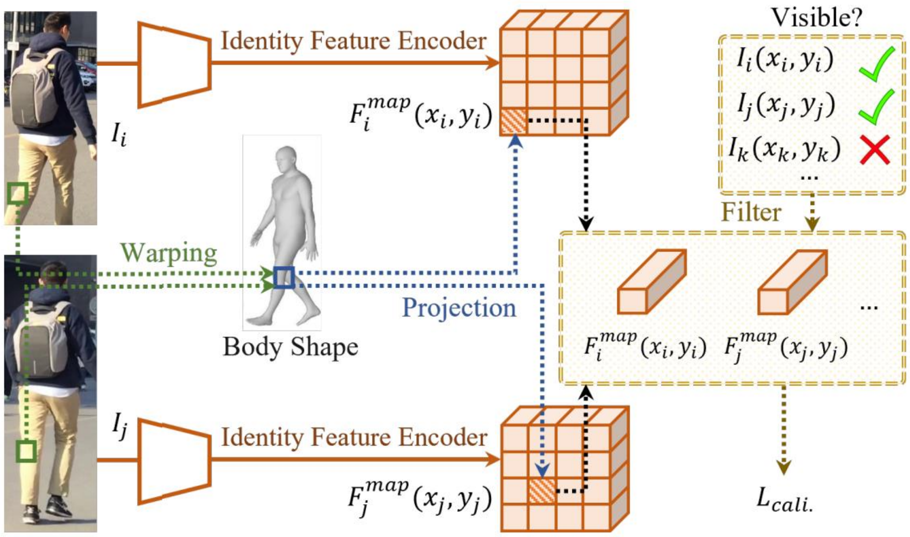
SEAS pipeline – using body shape as supervision

Saito, Shunsuke, et al. "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

Using 3-D Body Shape as Supervision (Calibration)

- Warping the point with a shared body shape model
- Sample the features from the corresponding area and determine their visibility
- Minimize the difference of features across frames

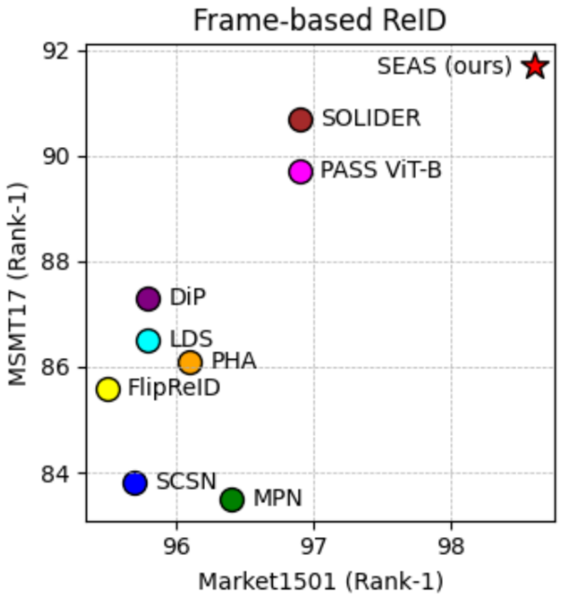
$$\mathcal{L}_{cali.} = \frac{1}{kn}\sum_{k}\sum_{n}\text{Variance}(\boldsymbol{F}_i^{map}(x_i, y_i))$$



USC

Main Result

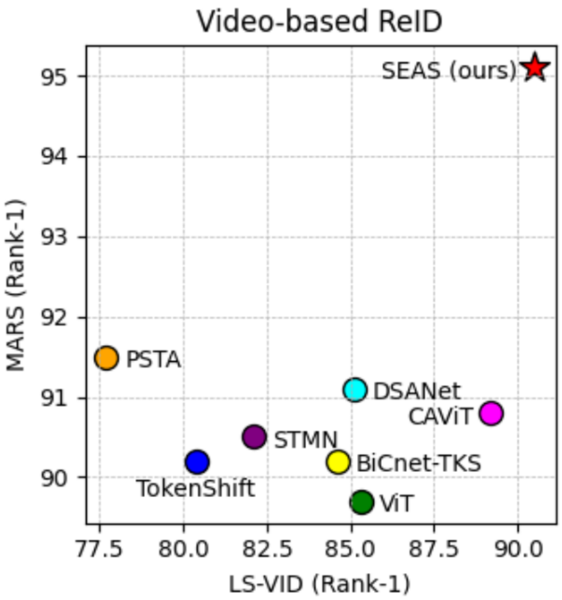
- Frame-based results



Method	Market1501 [81]		MSMT17 [68]	
	Rank-1	mAP	Rank-1	mAP
ViT-B [13]	94.0	87.6	82.8	63.6
TransReID [22]	95.2	89.5	86.2	69.4
AAFormer [90]	95.4	87.7	83.6	63.2
AGW [75]	95.5	89.5	81.2	59.7
FlipReID [47]	95.5	89.6	85.6	68.0
CAL [53]	95.5	89.5	84.2	64.0
PFD [63]	95.5	89.7	83.8	64.4
SAN [30]	96.1	88.0	79.2	55.7
LDS [76]	95.8	90.4	86.5	67.2
DiP [36]	95.8	90.3	87.3	71.8
MPN [12]	96.4	90.1	83.5	62.7
MSINet [18]	95.3	89.6	81.0	59.6
SCSN [9]	95.7	88.5	83.8	58.5
PHA [77]	96.1	90.2	86.1	68.9
PASS ViT-B [89]	<u>96.9</u>	93.3	89.7	74.3
SOLIDER [8]	<u>96.9</u>	<u>93.9</u>	<u>90.7</u>	<u>77.1</u>
ASSP* [5]	95.0	87.3	-	-
3DInvarReID* [40]	95.1	87.9	80.8	59.1
Baseline (ResNet-50)	94.1	83.2	73.8	47.2
SEAS (ResNet-50)	**98.6**	**98.9**	**91.7**	**92.8**

Main Result

❑ Video-based results



Method	MARS [82]		LS-VID [37]	
	Rank-1	mAP	Rank-1	mAP
GRL [42]	91.0	84.8	-	-
TokenShift [4]	90.2	86.6	80.4	68.7
ViT [13]	89.7	86.4	85.3	76.4
TCLNet [26]	89.8	85.1	-	-
AP3D [19]	90.1	85.1	-	-
DenseIL [23]	90.8	87.0	-	-
STMN [15]	90.5	84.5	82.1	69.2
BiCnet-TKS [27]	90.2	86.0	84.6	75.1
STRF [1]	90.3	86.1	-	-
RFCnet [28]	90.7	86.3	-	-
CTL [41]	91.4	86.7	-	-
DSANet [32]	91.1	86.6	85.1	75.5
CAViT [72]	90.8	87.2	89.2	79.2
Baseline (PSTA) [65]	91.5	85.8	77.7	67.2
SEAS (PSTA)	**95.1**	**96.6**	**90.5**	**93.4**

Wang, Yingquan, et al. "Pyramid spatial-temporal aggregation for video-based person re-identification." *ICCV*. 2021.

Ablation Results

	Method	Rank-1	mAP	Params	FLOPs
(I) Appearance	Baseline (Market1501)	94.1	83.2	23.51M	4.07G
(II) Body shape as input	+ PIFu as 2^{nd} branch	94.1 (+0.0)	84.8 (+1.6)	34.80M	6.28G
	+ PIFu concatenation	94.3 (+0.2)	85.8 (+2.6)	34.89M	4.26G

Including 3-D body shape as input slightly improves the mAP, while rank-1 accuracy does not show too much difference.

Ablation Results

	Method	Rank-1	mAP	Params	FLOPs
(I) Appearance	Baseline (Market1501)	94.1	83.2	23.51M	4.07G
(II) Body shape as input	+ PIFu as 2^{nd} branch	94.1 (+0.0)	84.8 (+1.6)	34.80M	6.28G
	+ PIFu concatenation	94.3 (+0.2)	85.8 (+2.6)	34.89M	4.26G
(III) Body shape as supervision	+ SEAS (SPIN)	97.1 (+3.0)	97.8 (+14.6)	23.51M	4.07G
	+ SEAS (PIFu)	**98.6** (+4.5)	**98.9** (+15.7)	23.51M	4.07G

Using SEAS for 3-D body shape supervision significantly boosts the performance without introducing extra computation cost.

Ablation Results

	Method	Rank-1	mAP	Params	FLOPs
(I) Appearance	Baseline (Market1501)	94.1	83.2	23.51M	4.07G
(II) Body shape as input	+ PIFu as 2^{nd} branch	94.1 (+0.0)	84.8 (+1.6)	34.80M	6.28G
	+ PIFu concatenation	94.3 (+0.2)	85.8 (+2.6)	34.89M	4.26G
(III) Body shape as supervision	+ SEAS (SPIN)	97.1 (+3.0)	97.8 (+14.6)	23.51M	4.07G
	+ SEAS (PIFu)	**98.6** (+4.5)	**98.9** (+15.7)	23.51M	4.07G
(IV) SEAS w/ calibration for video frames	Baseline (MARS)	91.5	85.8	35.43M	37.70G
	+ SEAS (w/o $\mathcal{L}_{cali.}$)	94.8 (+3.3)	96.5 (+10.7)	35.43M	37.70G
	+ SEAS (w/ $\mathcal{L}_{cali.}$)	**95.1** (+3.6)	**96.7** (+10.9)	35.43M	37.70G

SEAS assists the video-based recognition, and the calibration across different frames further improves the performance slightly.

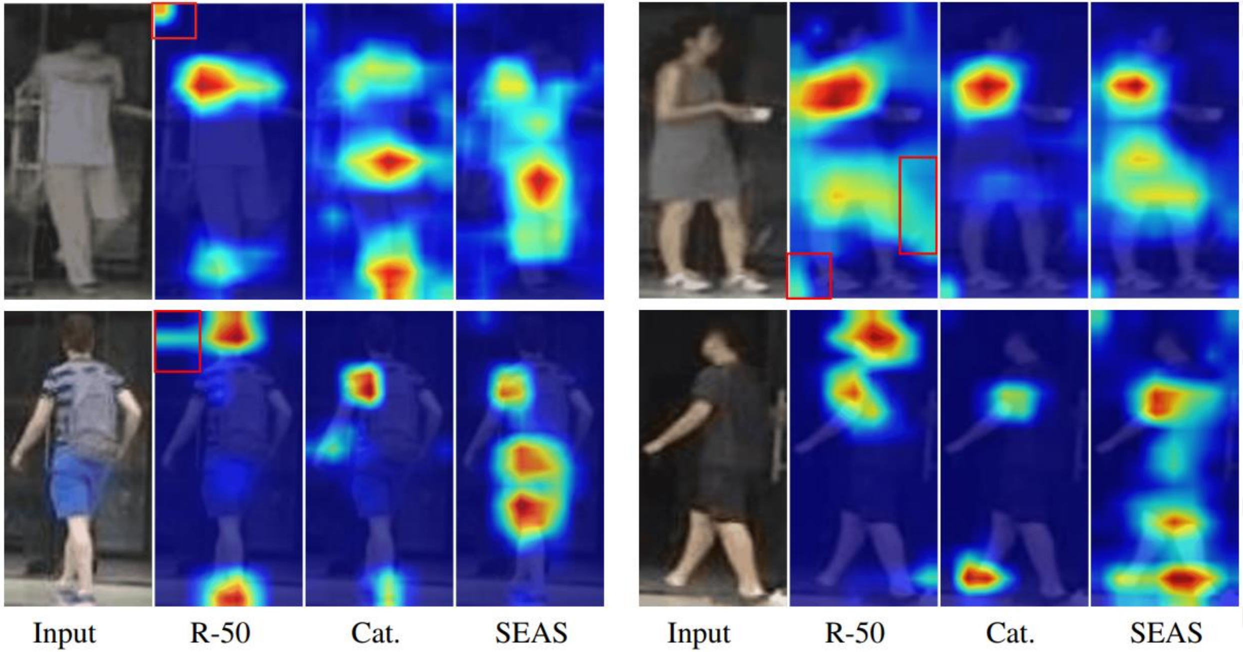
Ablation Studies on Generalizability

Method	Market1501 [81]		MSMT17 [68]	
	Rank-1	mAP	Rank-1	mAP
BoT [44]	94.5	85.9	74.1	50.2
w/ SEAS	**95.9**	**97.5**	**81.3**	**86.2**
LDS [76]	95.8	90.4	86.5	67.2
w/ SEAS	**96.3**	**97.8**	**86.6**	**90.1**

Method	MARS [82]		LS-VID [37]	
	Rank-1	mAP	Rank-1	mAP
STMN [15]	90.5	84.5	82.1	69.2
w/ SEAS	**92.2**	**94.9**	**84.1**	**88.9**
BiCnet-TKS [27]	**90.2**	86.0	84.6	75.1
w/ SEAS	90.1	**87.9**	**86.7**	**90.8**

SEAS can also be applied to other architectures for both frame-based and video-based re-identification tasks.

Ablation Studies on Generalizability



Compared with concatenation of features, SEAS force the attention to distribute more evenly across different body parts

CaesarNeRF: Calibrated Sematic Representation for Few-shot Generalizable Neural Rendering

Under Review

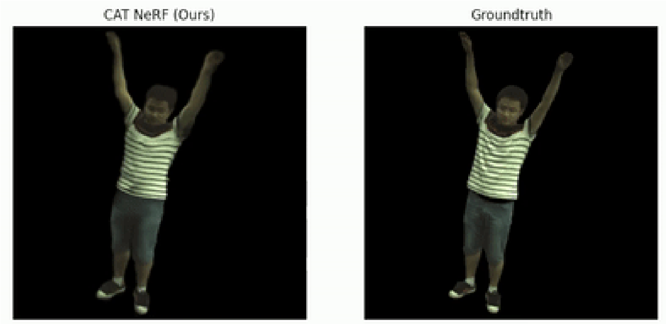
Haidong Zhu[1*], Tianyu Ding[2*], Tianyi Chen[2],
Ilya Zharkov[2], Ram Nevatia[1], Luming Liang[2]

University of Southern California[1] Microsoft[2]

NeRF for Body Shape Representation

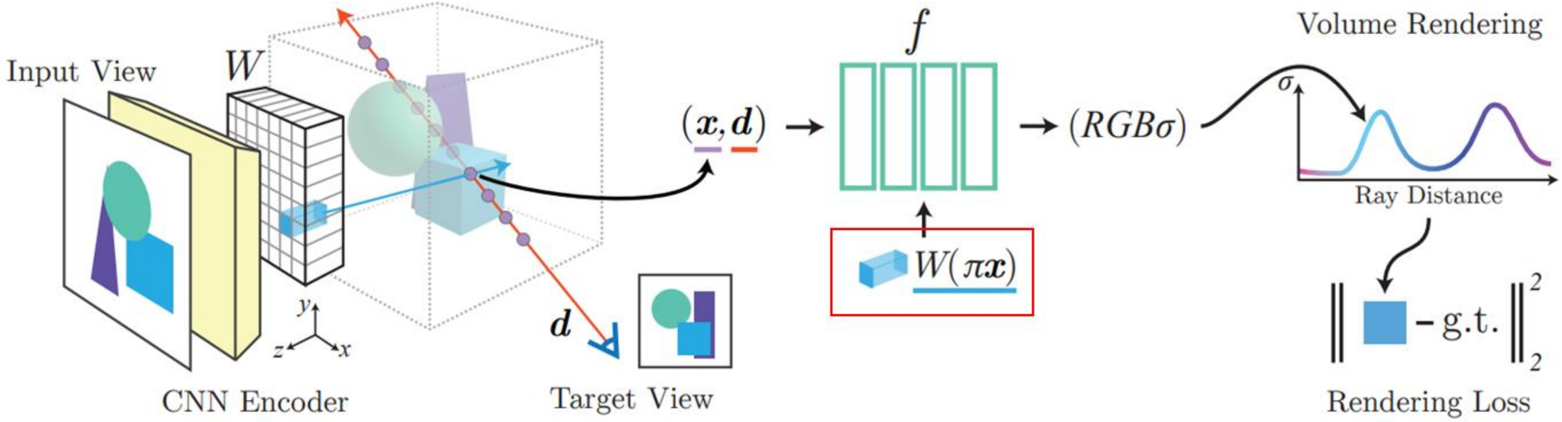
- Rendering results from recent NeRF models are of better quality than PIFu
- NeRF does not require 3-D shape for training
- Requirement to use a NeRF model for Re-ID
  - **Generalizability;**
  - **Number of Reference Views Required**.

Source image, SMPL, PIFu



NeRF Rendered results

NeRF and Generalizable NeRF



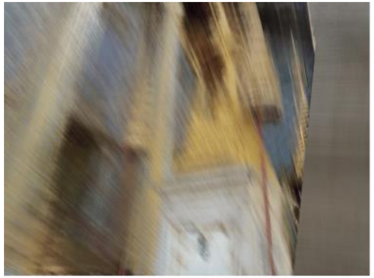
- Render an image from novel viewpoint with given images
- (x, y, z, ray_direction, <u>Pixel level embeddings</u>) → RGB + density

Image source: Yu, Alex, et al. PixelNeRF. *CVPR*. 2021.

NeRF and Generalizable NeRF

PSNR on LLFF

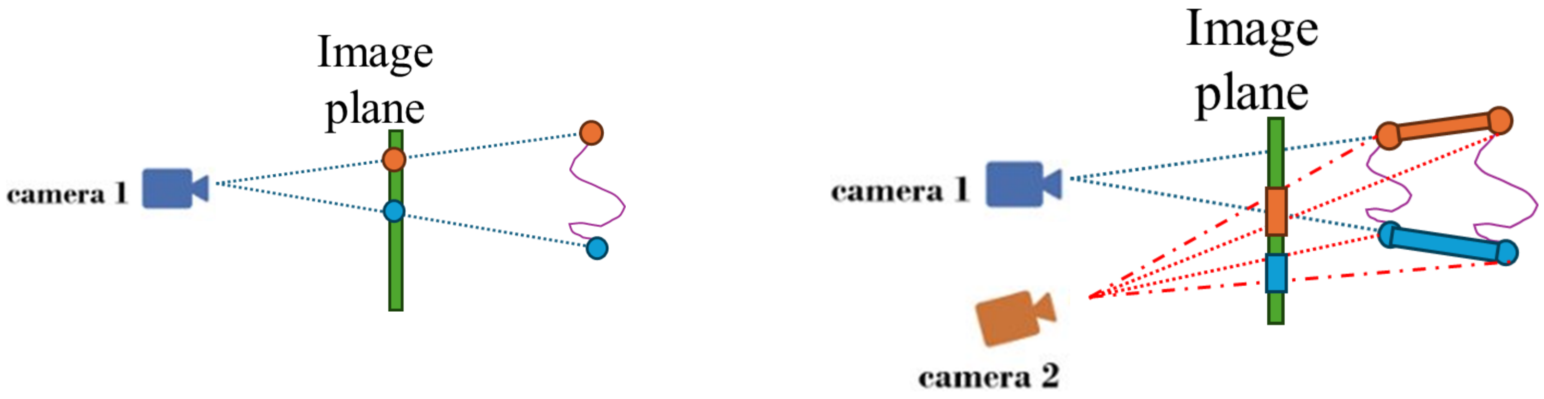
# of views	GeoNeRF	GNT
>=10	25.44	25.65
2	18.76	20.88
1	-	16.57



- Reasonable performance with all (>10) reference views
- Performance dropped significantly with limited number of input views

Johari M.M., et al. GeoNeRF. *CVPR 2022*
Varma M. et al. GNT. *ICLR 2023*

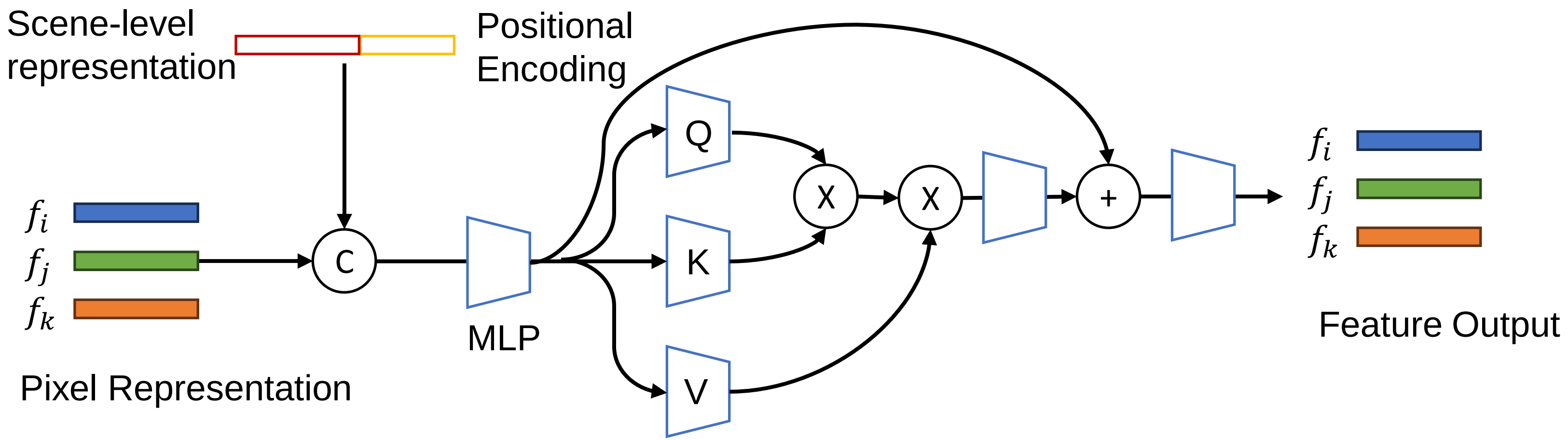
Few-shot Generalizable NeRF



Generalizable rendering with limited reference views:

- ➢ Relation ambiguity between different points at different camera pose;
- ➢ Pixel-level projection only intakes one pixel feature without a scene-level understanding.

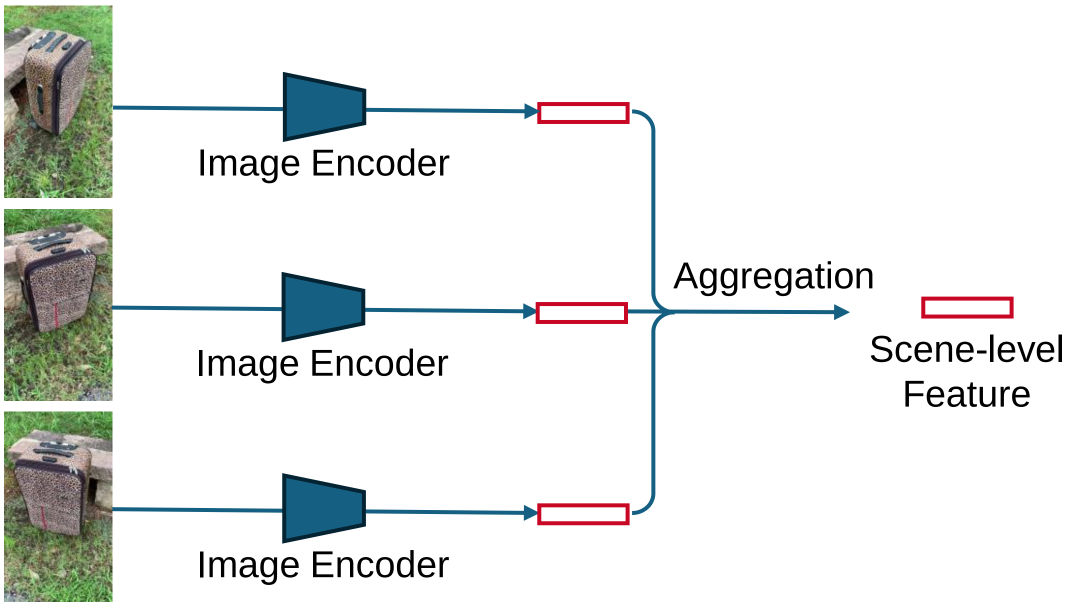
Scene-level Representation



Scene-level latent representation

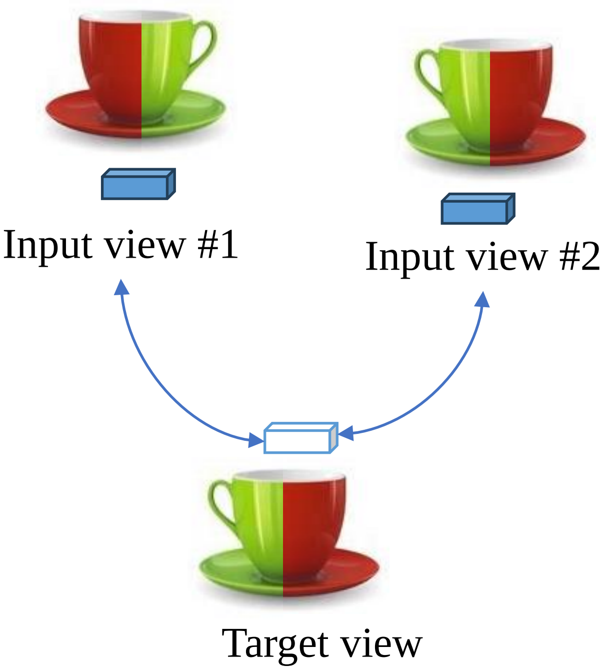
- ➢ Shared with different points of the same scene
- ➢ Concatenate with pixel-level embeddings
- ➢ Encoded with self-attention for corresponding feature encoding.

Scene-level Representation



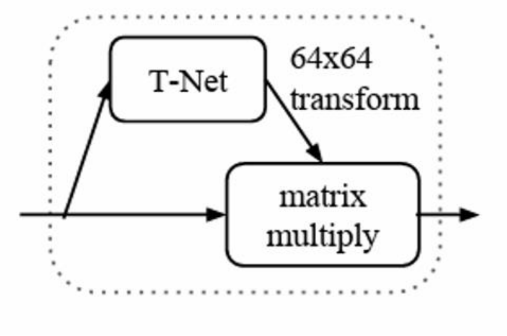
Calibration Across Different Views

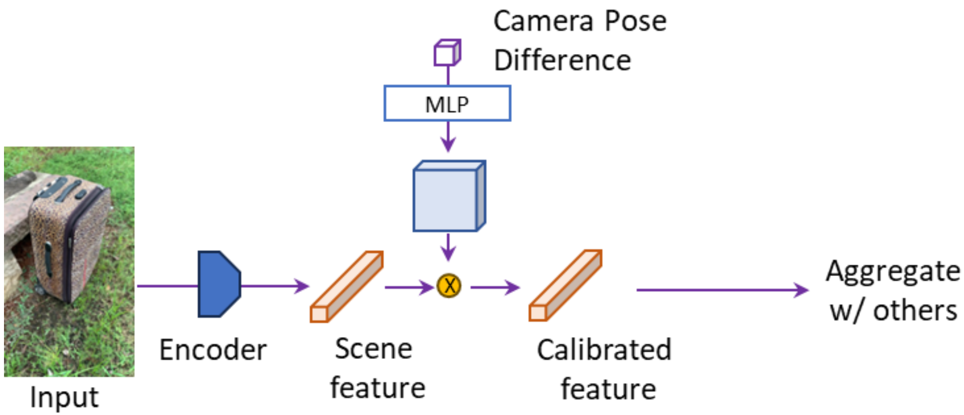
- Scene-level features do not include explicit camera poses from different views;
- Features encoded from different views suffer from conflict between them;
- Our target has only one view, and transformations between input and target are available.



Calibration Across Different Views

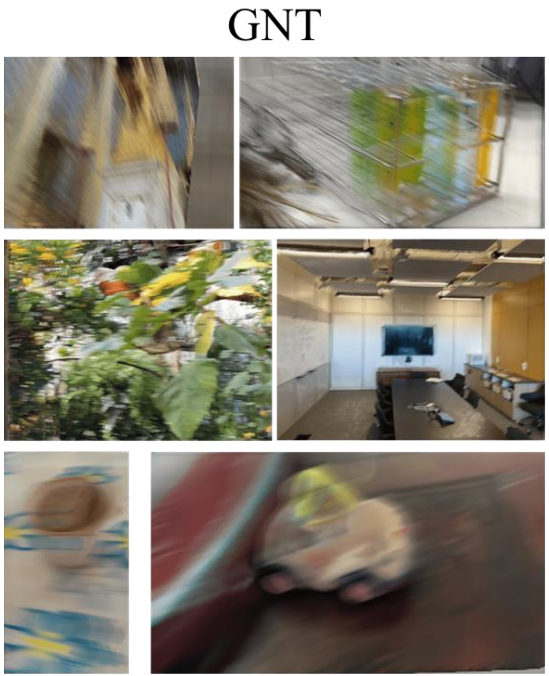
We encode the camera pose by converting it to a transformation matric in the feature space and multiply it with the scene feature for calibration.

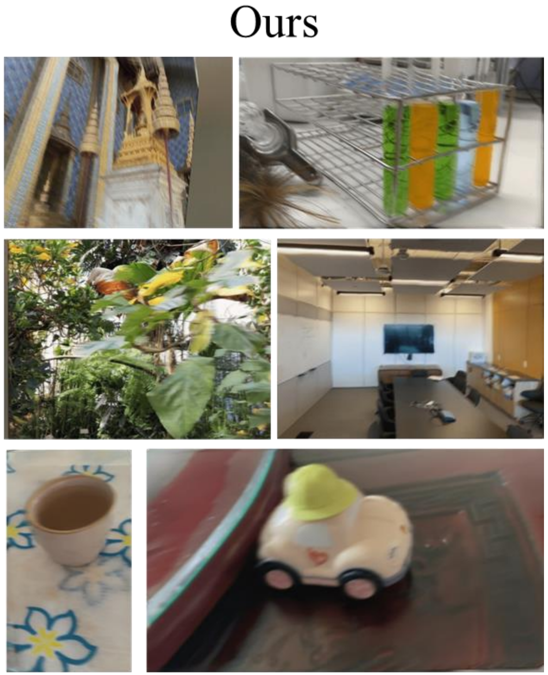




Qi, Charles R., et al. PointNet. *CVPR* 2017.

Main Results

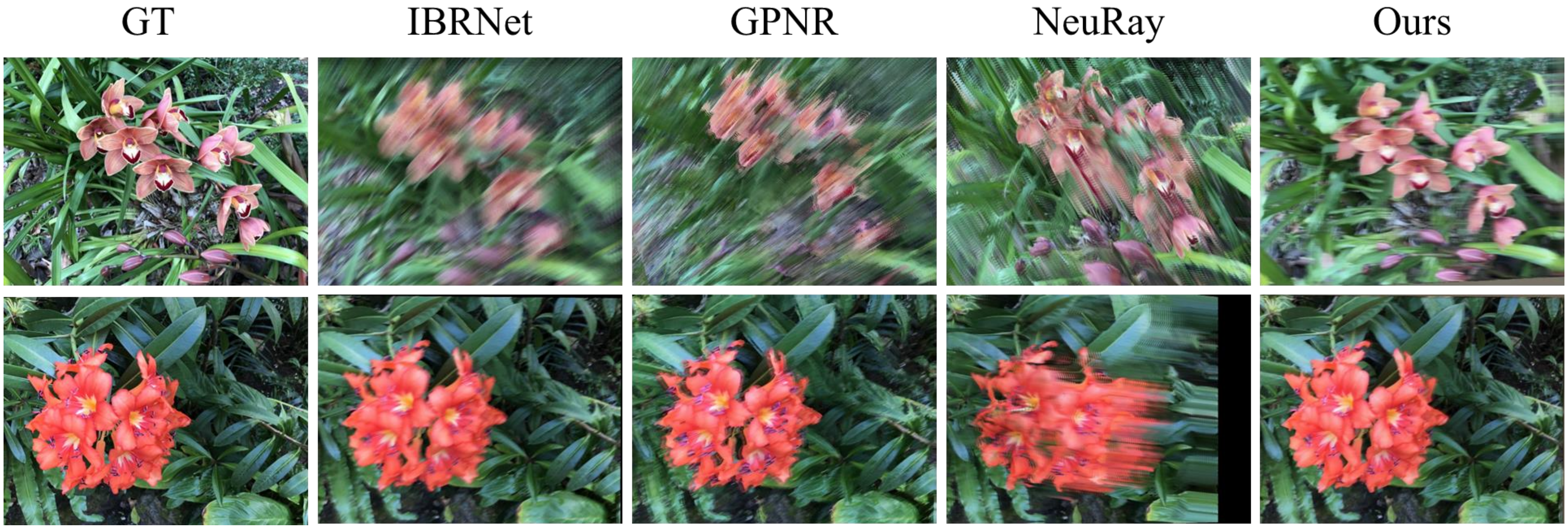




Rendered results with one view as reference, compared with our baseline, GNT

Wang, Peihao, et al. "Is Attention All NeRF Needs?." *ICLR 2023*.

Main Results



Rendered results with one view as reference, compared with other SOTA methods

Main Results

Rendered results with two views as reference, compared with our baseline, GNT

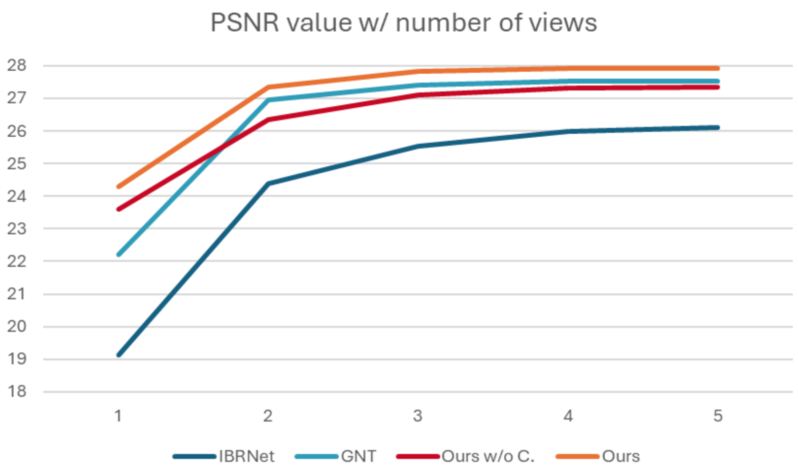
Wang, Peihao, et al. "Is Attention All NeRF Needs?." *ICLR 2023*.

Main Results

Method	1 reference view			2 reference views			3 reference views		
	PSNR (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)	SSIM (↑)
PixelNeRF [76]	9.32	0.898	0.264	11.23	0.766	0.282	11.24	0.671	0.486
GPNR [58]	15.91	0.527	0.400	18.79	0.380	0.575	21.57	0.288	0.695
NeuRay [38]	16.18	0.584	0.393	17.71	0.336	0.646	18.26	0.310	0.672
GeoNeRF [25]	-	-	-	18.76	0.473	0.500	23.40	0.246	0.766
MatchNeRF [8]	-	-	-	21.08	0.272	0.689	22.30	0.234	0.731
MVSNeRF [6]	-	-	-	19.15	0.336	0.704	19.84	0.314	0.729
IBRNet [64]	16.85	0.542	0.507	21.25	0.333	0.685	23.00	0.262	0.752
GNT [60]	16.57	0.500	0.424	20.88	0.251	0.691	23.21	0.178	0.782
Ours	18.31	0.435	0.521	21.94	0.224	0.736	23.45	0.176	0.794

Results with 1,2, and 3 views as reference on the LLFF dataset

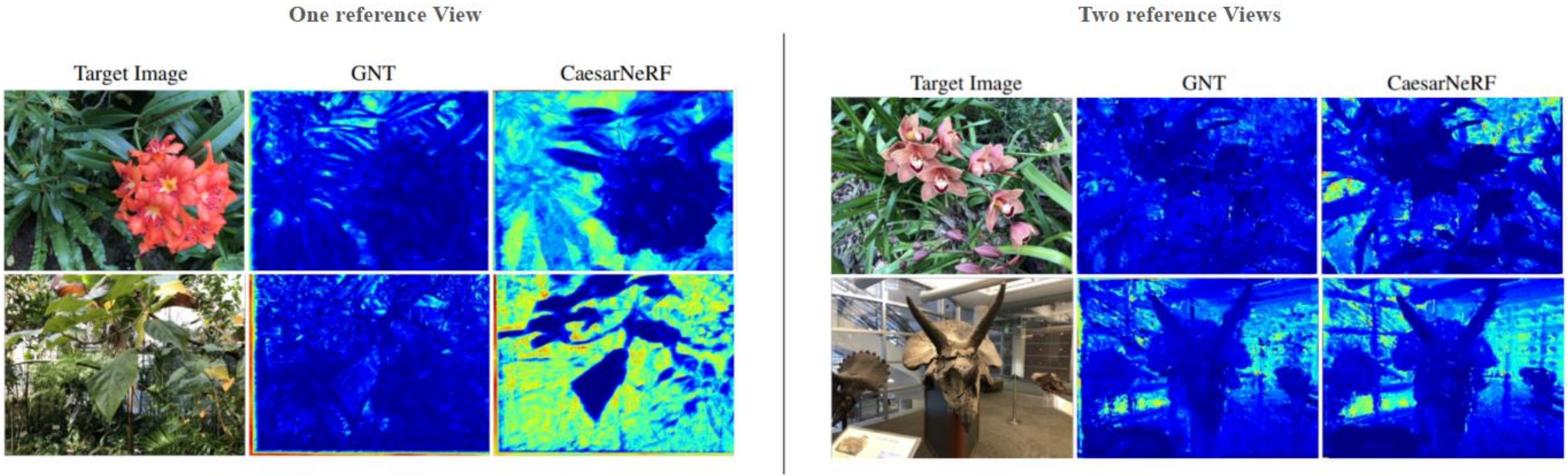
Main Results



		PSNR	LPIPS	SSIM
1-view	IBRNet	19.14	0.458	0.595
	GNT	22.22	0.433	0.678
	Ours *w/o* C.	23.61	0.371	0.718
	Ours	24.28	0.334	0.747
2-view	IBRNet	24.38	0.266	0.818
	GNT	26.94	0.236	0.850
	Ours *w/o* C.	26.34	0.274	0.817
	Ours	27.34	0.215	0.856
3-view	IBRNet	25.53	0.203	0.858
	GNT	27.41	0.206	0.870
	Ours *w/o* C.	27.10	0.228	0.850
	Ours	27.82	0.190	0.875
4-view	IBRNet	25.99	0.190	0.867
	GNT	27.51	0.197	0.875
	Ours *w/o* C.	27.30	0.210	0.862
	Ours	27.92	0.181	0.881
5-view	IBRNet	26.12	0.188	0.867
	GNT	27.51	0.194	0.876
	Ours *w/o* C.	27.34	0.203	0.865
	Ours	27.92	0.179	0.882

Results with comparing using calibration with no calibration on MVImgNet

Main Result



Depth Prediction using one or two reference views, comparing with GNT

Wang, Peihao, et al. "Is Attention All NeRF Needs?." *ICLR 2023*.

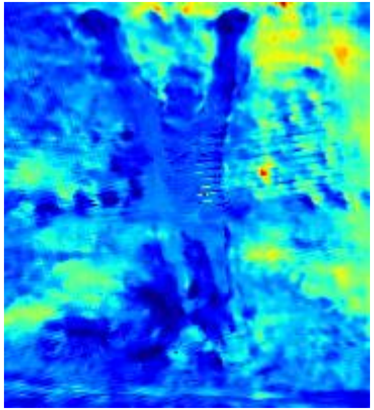
Human Result

Input # 1

Input # 2

Input # 3

Use all input views



Depth map using Input #1

Human Result

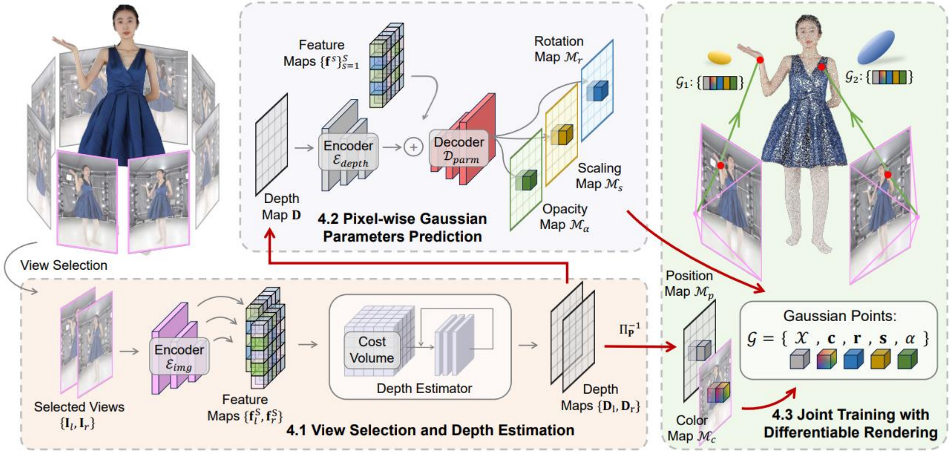
	Rank-1	mAP
Baseline (ResNet 50)	73.8	47.2
+ SEAS (PIFu)	**91.7**	92.8
+ SEAS (CeasarNeRF)	91.4	**93.0**

Using the encoder of CaesarNeRF as 3-D shape extractor with SEAS on MSMT17 dataset

Summary

- Multimodal features enable recognition at long distances in the wild, which includes body shape, appearance and specific activities;
- Compared with using 3-D body shape as input, using it as supervision can provide more distinguishment for encoded features;
- By using calibrated semantic representation, we can extend generalizable NeRF with limited reference views, but a good feature encoder still requires additional human body priors.

Future Directions



Generalizable Few-shot Rendering with Human Prior

- Depth map for more accurate human object guidance;
- Generalizable body-specific models.

Zheng, Shunyuan, et al. "Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis." arXiv preprint arXiv:2312.02155 (2023).

Future Directions

Original image

Generated image

Vision-language model assisted person re-identification

- ➢ Prompts for clothes changes with shape prior
- ➢ Augment model training with different clothes variations.

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022

Publications

- 3-D Reconstruction
  - Zhu* et al., CaesarNeRF: Calibrated Semantic Representation for Few-shot Generalizable Neural Rendering, **under review**
  - Zhu et al., Multimodal neural radiance field, **ICRA 2023**
  - Zhu et al., CAT-NeRF: Constancy-Aware Tx2Former for Dynamic Body Modeling, **CVPRw 2023**
  - Zhu, et al., Open: Order-preserving pointcloud encoder decoder network for body shape refinement, **ICPR 2022**
  - Duan*, Zhu* et al. Curriculum deepsdf, **ECCV 2020**
- Re-Identification
  - Zhu et al., SEAS: Shape-aligned supervision for person re-identification, **CVPR 2024**.
  - Zhu et al., Sharc: Shape and appearance recognition for person identification in-the-wild, **WACV 2024**.
  - Zheng*, Zhu* et al. GaitSTR: Gait Recognition with Sequential Two-stream Refinement, **TBIOM 2024**.
  - Zhu* et al. Gaitref: Gait recognition with refined sequential skeletons, **IJCB 2023.**
  - Zhu et al. Gait recognition using 3-d human body shape inference, **WACV 2023.**
  - Zhu et al., Temporal shift and attention modules for graphical skeleton action recognition, **ICPR 2022.**

Publications

- Vision and Language
  - Zheng et al., Large Language Models are Good Prompt Learners for Low-Shot Image Classification, **CVPR 2024**.
  - Zheng, Zhu et al., CAILA: Concept-Aware Intra-Layer Adapters for Compositional Zero-Shot Learning, **WACV 2024**.
  - Zhu et al., Self-supervised Learning for Sentiment Analysis via Image-text Matching, **ICASSP 2022**.
  - Zhu, et al., Utilizing Every Image Object for Semi-supervised Phrase Grounding, **WACV 2021.**
  - He, Zhu, et al., CPARR: Category-based Proposal Analysis for Referring Relationships, **CVPRw 2020.**
- System and Survey
  - Ding et al., The efficiency spectrum of large language models: An algorithmic survey, **arXiv 2023**.
  - Nguyen et al., AG-ReID 2023: Aerial-Ground Person Re-identification Challenge Results, **IJCB 2023**.
  - Li et al., GAIA at SMKBP 2020-a dockerlized multi-media multi-lingual knowledge extraction, clustering, temporal tracking and hypothesis generation system, **TAC 2020**.

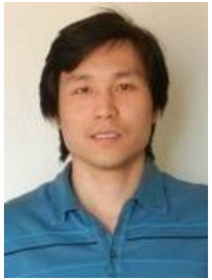
Acknowledgement

Pranav Budhwant

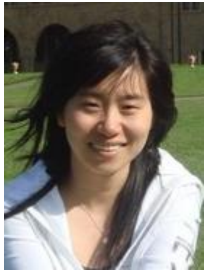
Tianyu Ding

Yueqi Duan

Xuefeng Hu



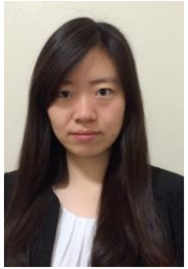
Luming Liang



Jiajia Luo

Ram Nevatia

Arka Sadhu



Ye Yuan

Wanrong Zheng

Zhaoheng Zheng

And many others…

Thank you!