HiFace: High-Fidelity 3D Face Reconstruction by Learning Static and Dynamic Details

Experiments

Ablation Studies

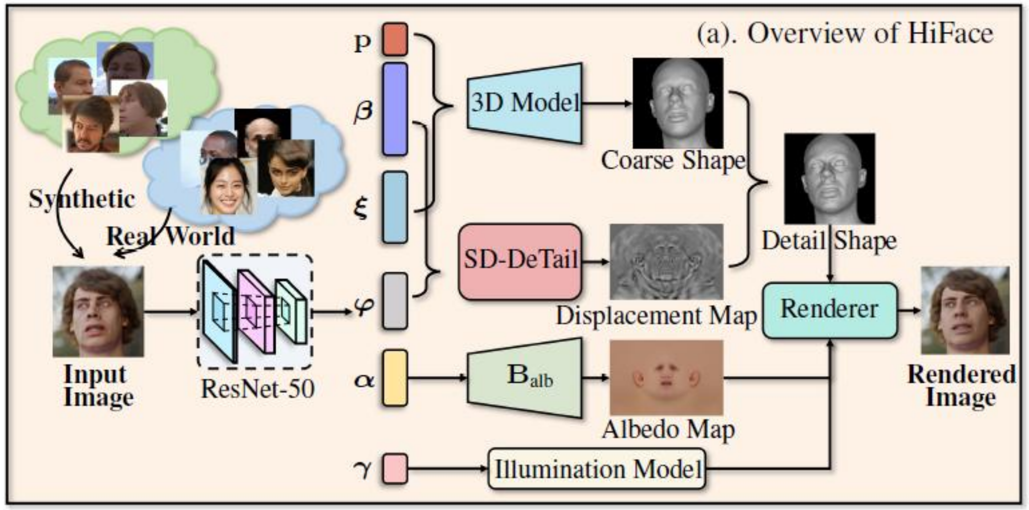
Quick Review

Limitations of Existing Methods

- Traditional 3DMM-based methods **fail to separate static and dynamic details.**
- For example, wrinkles from an old person's face might be unnaturally transferred to a young person.
- Existing approaches use **image-level supervision only**, leading to poor decoupling of static and dynamic details.

Key Idea of Hiface

- Static Detail (Person-Specific Feature): Uses **PCA-based Displacement Basis** to capture identity-specific facial details
- Dynamic Detail (Expression-Based Wrinkles): Modeled through **interpolation between compressed and stretched displacement maps**

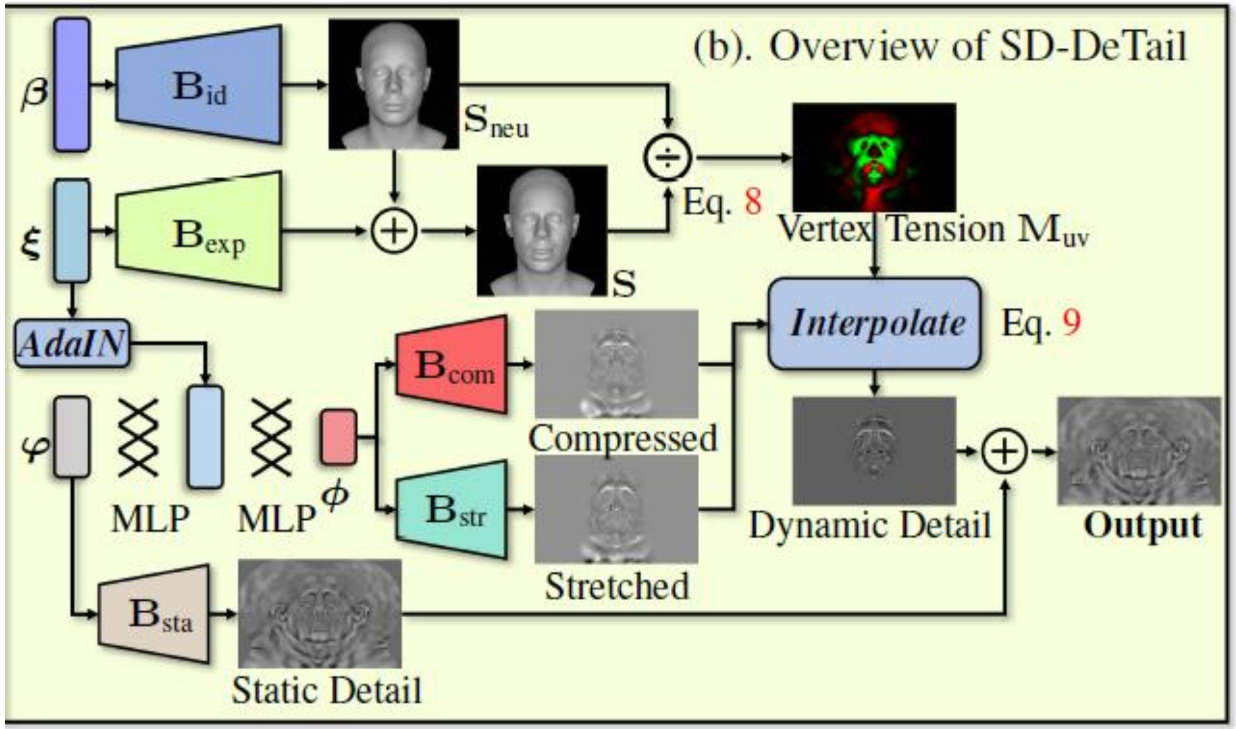


p: pose coef β: identity coef ξ: expression coef φ: static coef α: albedo coef γ: SH coef ϕ: dynamic coef

Key Idea of Hiface

- SD-DeTail Module:

Separates and combines static & dynamic details in one module



p: pose coef β: identity coef ξ: expression coef φ: static coef α: albedo coef γ: SH coef ϕ: dynamic coef

$D_d = M \odot D_c + (1 - M) \odot D_s$

Learning framework

Datasets and loss function

Dataset

- Hybrid Dataset
  - Synthetic dataset
  - Real-world dataset
- Synthetic dataset
  - generated in synthetic dataset pipeline
  - **has GT labels**
  - Ground truth vertices, landmarks, albedo, displacement maps
- Real-world dataset **for generalization in wild**
  - use pre-trained dense landmark detector
  - no labels->**Self supervised learning loss functions**

Loss Functions

- Detail Losses (use Ground-Truth from Synthetic dataset)
- displacement maps (height displacement in UV map)

$$\mathcal{L}_{\text{sta}} = \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{sta}} - \hat{\mathbf{D}}_{\text{sta}})\right\|_2$$
$$\mathcal{L}_{\text{com}} = \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{com}} - \hat{\mathbf{D}}_{\text{com}})\right\|_2,$$
$$\mathcal{L}_{\text{str}} = \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{str}} - \hat{\mathbf{D}}_{\text{str}})\right\|_2$$
$$\mathcal{L}_{\text{detail}} = \mathcal{L}_{\text{sta}} + \mathcal{L}_{\text{com}} + \mathcal{L}_{\text{str}}$$

Loss Functions

- Coarse shape Losses (use GT + KL divergence)
- vertices
- KL divergence loss for overfitting

$$\mathcal{L}_{\text{ver}} = \left\| \mathbf{M}_{\text{ver}} \odot (\mathbf{S} - \hat{\mathbf{S}}) \right\|_2, \qquad (11)$$

$$\mathcal{L}_{\text{kl}} = \rho(\beta) \big(\log \rho(\beta) - \log \rho(\hat{\beta}) \big), \qquad (12)$$

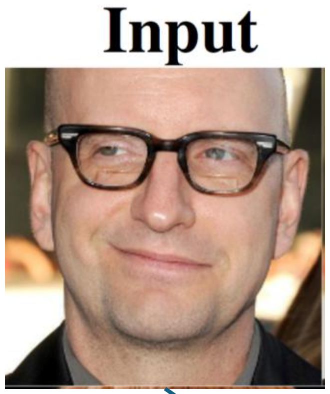
$$\mathcal{L}_{\text{shp}} = \mathcal{L}_{\text{ver}} + \mathcal{L}_{\text{kl}}.$$

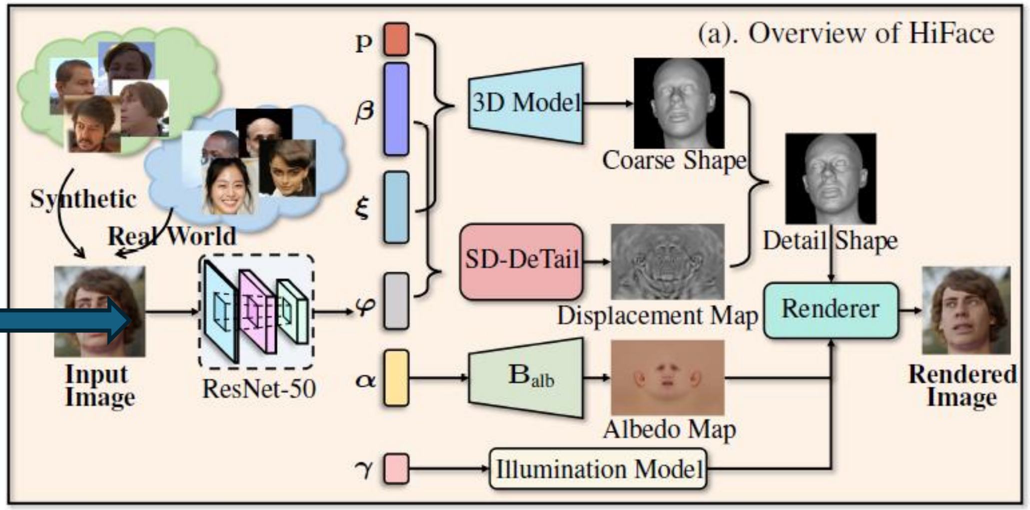
Loss Functions

$$\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{pho}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}}, \tag{13}$$

- **Self-supervised losses** (used for Real-world dataset)

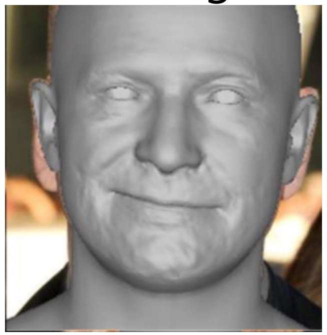
Real-world 2D image





3D model

2D Image



reduce differences
(in 3 different features)

Loss function

$$\mathcal{L}_{\text{kd}} = \mathbf{\Gamma}_{\text{age}}(\mathbf{I})\left(\log \mathbf{\Gamma}_{\text{age}}(\mathbf{I}) - \log \hat{\mathbf{p}}_{\text{age}}\right).$$

- **knowledge distillation loss** (지식 증류)
- consider the static detail **heavily correlates to person specific age attribute**
- static detail coefficient => get P_{age}
- P_{age}: age classification probabilities
- Γ_{age}: pre-trained age recognition model

make similar

φ $\xrightarrow{\text{MLP}}$ $\hat{\mathbf{p}}_{\text{age}}$ $\mathbf{\Gamma}_{\text{age}}(\mathbf{I})$

Overall loss function

$$\begin{aligned}\mathcal{L} = \lambda_{\text{detail}}\mathcal{L}_{\text{detail}} + \lambda_{\text{shp}}\mathcal{L}_{\text{shp}} \\ + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}},\end{aligned} \tag{15}$$

4. Experiments

Experiment Objectives

- validate **whether the model effectively decouples static and dynamic details**
- evaluate **whether HiFace outperforms existing models** in reconstructing high-resolution 3D faces with realistic details

Dataset

- Synthetic dataset with GT: **200K** pictures
- Real-world dataset for self-supervise: **400K** pictures
  - mask the hair and accessory
- split data by training and validation

Model Implementation

- Use PyTorch, use PyTorch3D's differentiable rasterizer to render
- Training Setup
  - 35 epoch
  - 8 x NVIDEA Tesla V100 GPU, batch size = 320
  - initialize ResNet-50 to pre-trained model on ImageNet
  - use **Adam** optimizer, initial learning rate = 1e-4
- Preprocessing: align and resize
- Loss weights

$$\lambda_{detail} = 10, \lambda_{shp} = 1, \lambda_{self} = 1, \lambda_{id} = 0.1, \lambda_{lmk} = 0.5, \lambda_{kd} = 1, \lambda_{reg} = 10^{-3}.$$

Quantitative Evaluation - REALY

- Use **REALY benchmark**
  - evaluates errors in different facial regions: nose, mouth, forehead, cheeks.
  - Metric: Normalized Mean Squared Error (NMSE)

- $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ 는 예측값과 실제값 간의 평균 제곱 오차입니다.
- σ^2 는 실제값의 분산(variance)입니다.

$$NMSE = \frac{MSE}{\cancel{\sigma^2}}$$

Quantitative Evaluation Result

Table 1. **Quantitative comparison of 3D face reconstruction methods on REALY benchmark**. "-c" and "-d" indicate coarse and detail shape, respectively. $@\mathcal{R}_N/@\mathcal{R}_M/@\mathcal{R}_F/@\mathcal{R}_C$/all indicate errors in nose/mouth/forehead/cheek/all regions. We highlight the best method for the two groups respectively. HiFace achieves the best reconstruction performance in the overall error by a large margin. Each component in HiFace contributes to a better reconstruction quality. The reconstructed details of HiFace further boost the quality while previous methods [24, 19] modeling details with only image-level supervision even deteriorate the reconstruction accuracy.

Group	Methods / e (mm)	*frontal-view*					*side-view*				
		$@\mathcal{R}_N$	$@\mathcal{R}_M$	$@\mathcal{R}_F$	$@\mathcal{R}_C$	all	$@\mathcal{R}_N$	$@\mathcal{R}_M$	$@\mathcal{R}_F$	$@\mathcal{R}_C$	all
Coarse	Deep3D [21]	1.719±0.354	**1.368±0.439**	2.015±0.449	1.528±0.501	1.657	1.749±0.343	1.411±0.395	2.074±0.486	1.528±0.517	1.691
	MGCNet [64]	1.771±0.380	1.417±0.409	2.268±0.503	1.639±0.650	1.774	1.827±0.383	**1.409±0.418**	2.248±0.508	1.665±0.644	1.787
	3DDFA-v2 [29]	1.903±0.517	1.597±0.478	2.447±0.647	1.757±0.642	1.926	1.883±0.499	1.642±0.501	2.465±0.622	1.781±0.636	1.943
	DECA-c [24]	1.694±0.355	2.516±0.839	2.394±0.576	1.479±0.535	2.010	1.903±1.050	2.472±1.079	2.423±0.720	1.630±1.135	2.107
	SADRNet [61]	1.791±0.542	1.591±0.488	2.413±0.537	1.856±0.701	1.913	1.771±0.521	1.560±0.462	2.490±0.566	2.010±0.715	1.958
	EMOCA-c [19]	1.868±0.387	2.679±1.112	2.426±0.641	1.438±0.501	2.103	1.867±0.554	2.636±1.284	2.448±0.708	1.548±0.590	2.125
	MICA [81]	1.585±0.325	3.478±1.204	2.374±0.683	**1.099±0.324**	2.134	1.525±0.322	3.567±1.212	2.379±0.675	**1.109±0.325**	2.145
	Ours-c (w/o Syn. Data)†	1.227±0.407	1.787±0.439	1.454±0.382	1.762±0.436	1.558	1.187±0.379	1.826±0.490	1.470±0.426	1.653±0.450	1.534
	Ours-c	**1.054±0.317**	1.461±0.430	**1.331±0.347**	1.342±0.384	**1.297**	**0.992±0.246**	1.505±0.454	**1.427±0.400**	1.439±0.429	**1.341**
Detail	DECA-d [24]	2.138±0.461	2.802±0.868	2.457±0.559	1.443±0.498	2.210	2.286±1.103	2.684±1.041	2.519±0.718	1.555±0.822	2.261
	EMOCA-d [19]	2.532±0.539	2.929±1.106	2.595±0.631	1.495±0.469	2.388	2.455±0.636	2.948±1.292	2.606±0.686	1.599±0.563	2.402
	HRN [42]	1.722±0.330	**1.357±0.523**	1.995±0.476	**1.072±0.333**	1.537	1.642±0.310	**1.285±0.528**	1.906±0.479	**1.038±0.322**	1.468
	Ours-d (w/o Syn. Data)†	1.465±0.557	1.790±0.425	1.528±0.373	1.618±0.362	1.600	1.422±0.537	1.849±0.473	1.530±0.414	1.572±0.399	1.594
	Ours-d (w/o static)*	1.055±0.290	1.469±0.415	1.336±0.337	1.319±0.374	1.295	1.004±0.233	1.491±0.437	1.418±0.392	1.418±0.415	1.332
	Ours-d (w/o dynamic)*	1.069±0.318	1.469±0.414	1.358±0.336	1.270±0.344	1.292	0.991±0.239	1.496±0.437	1.411±0.393	1.375±0.402	1.318
	Ours-d	**1.036±0.280**	1.450±0.413	**1.324±0.334**	1.291±0.362	**1.275**	**0.985±0.237**	1.489±0.436	**1.399±0.388**	1.360±0.395	**1.308**

† To align the dataset scale, w/o Syn. Data indicates we train the model without using the ground-truth labels from the synthetic dataset.

* To eliminate the bias of coarse shape in estimating the reconstruction error, we fix the coarse shape and train the details with/without static and dynamic factors for comparisons.

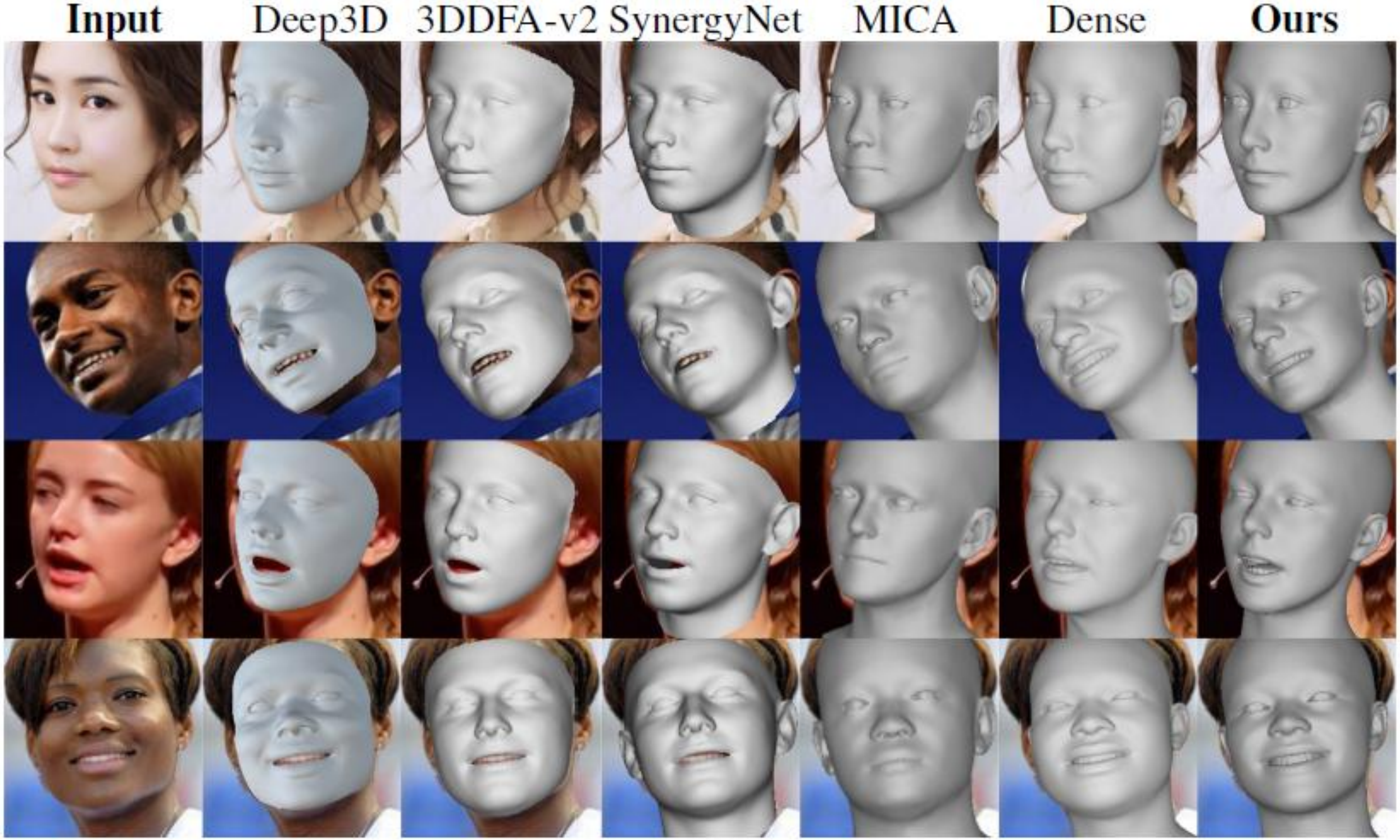
Quantitative Evaluation Result

- HiFace outperforms state-of-the-art (SOTA) methods by 15% in the REALY benchmark.
- Achieves lower reconstruction error across different face regions
- Compared to DECA and EMOCA (noisy), HiFace produces more natural facial expressions and details.
- Synthetic dataset is crucial also for detail
- HiFace effectively separates static and dynamic details while learning

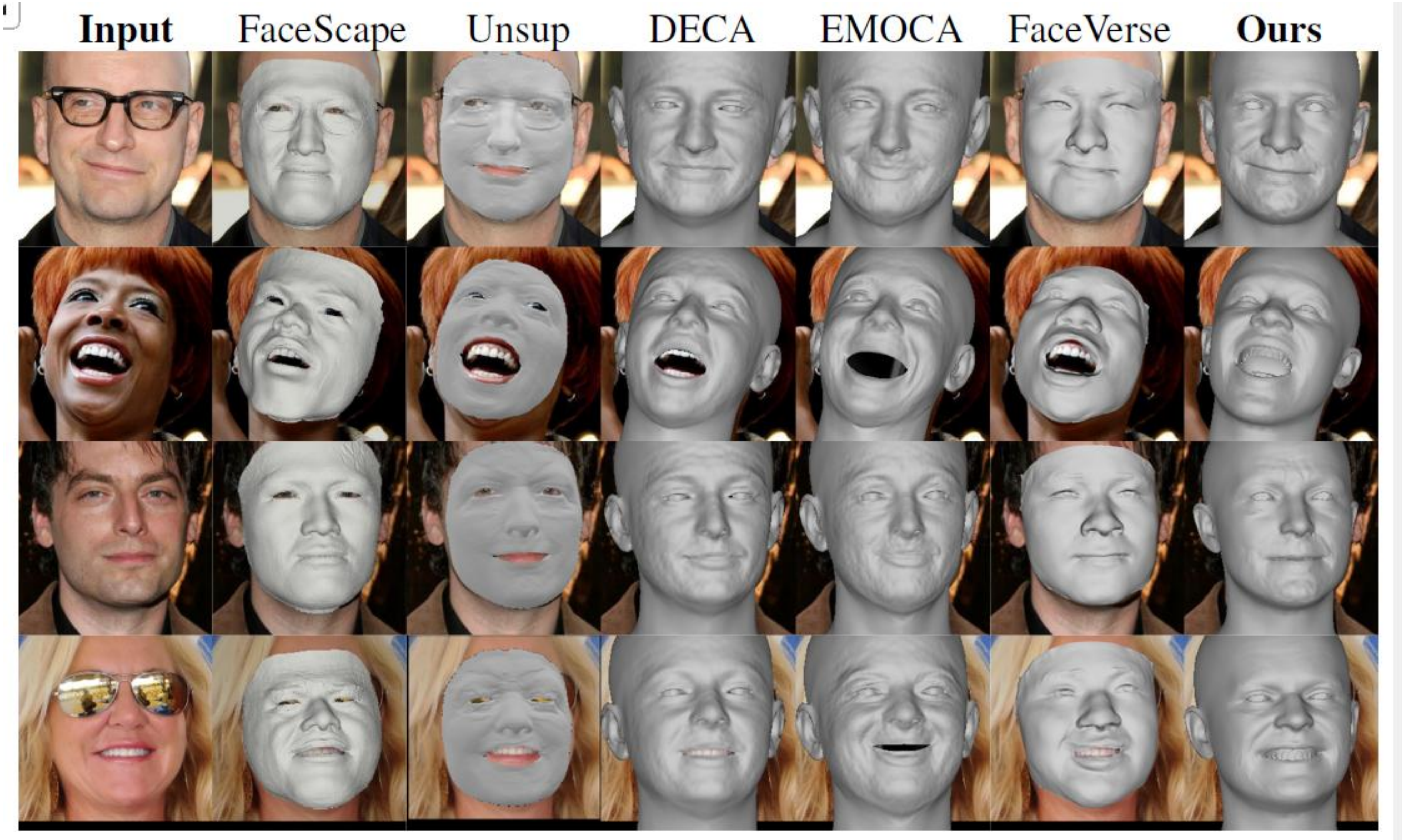
Qualitative Evaluation-visual comparison

- Visual comparisons were conducted against existing models (Deep3D, 3DDFA-v2, MGCNet, DECA, EMOCA, etc.).
    - Coarse
    - Detail
- HiFace was tested on **real-world face images** to assess **realism and detail accuracy**.

Qualitative Evaluation Result – Coarse Shape

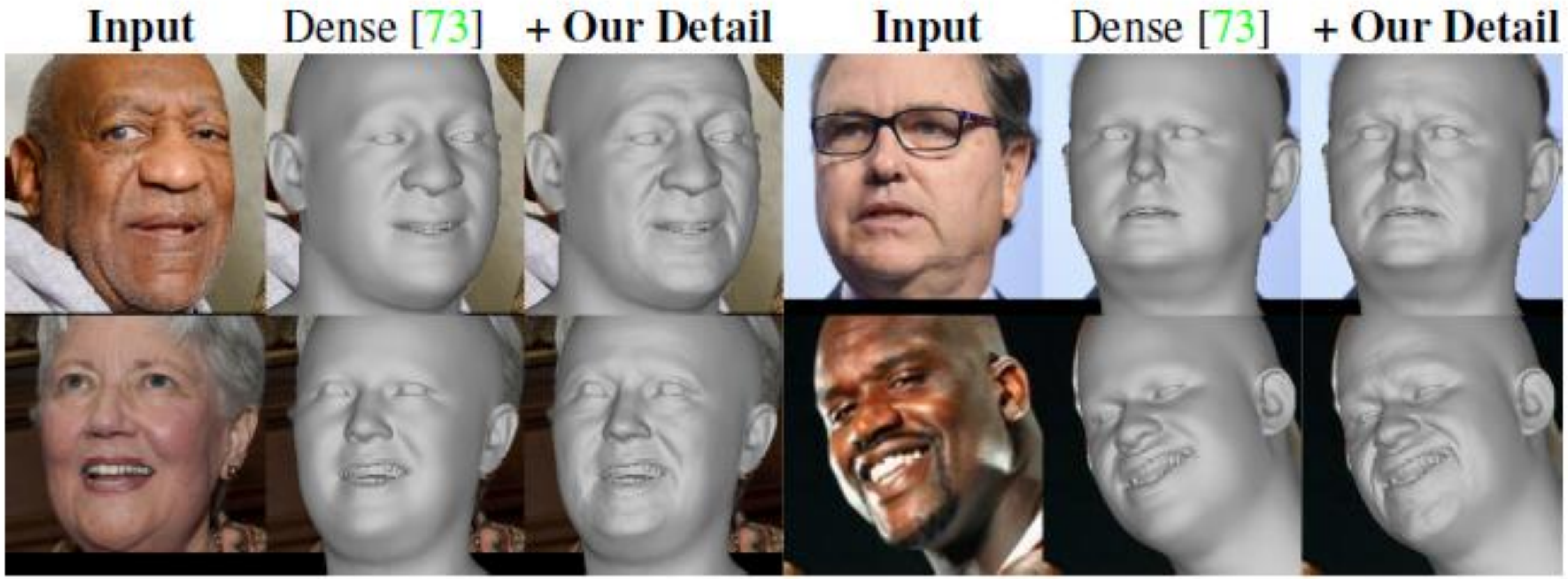


Qualitative Evaluation Result – With detail

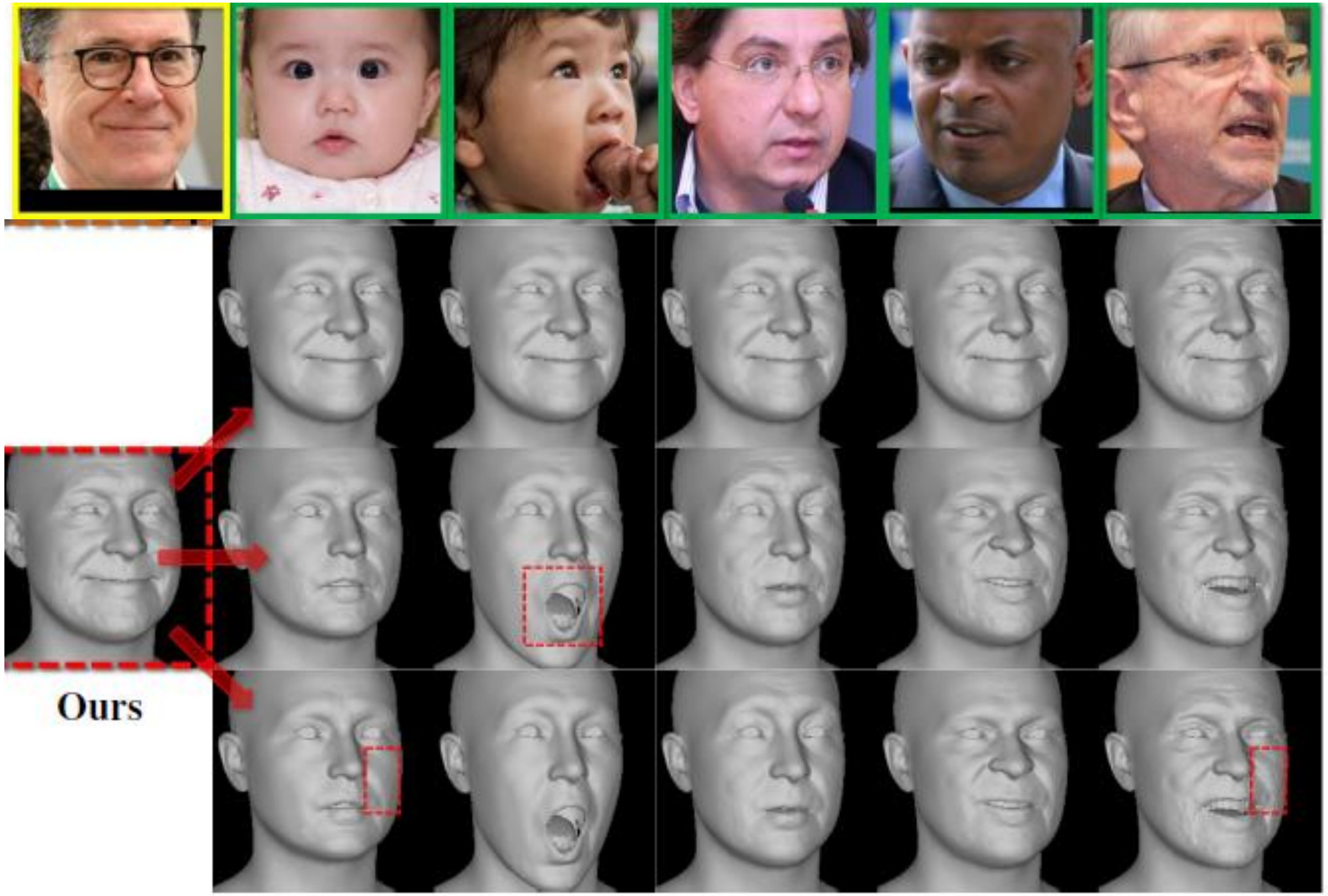


Dense + SD-DeTail

give identity, expression coefficient of Dense to SD-DeTail



Animated with Detail (static, dynamic, both)

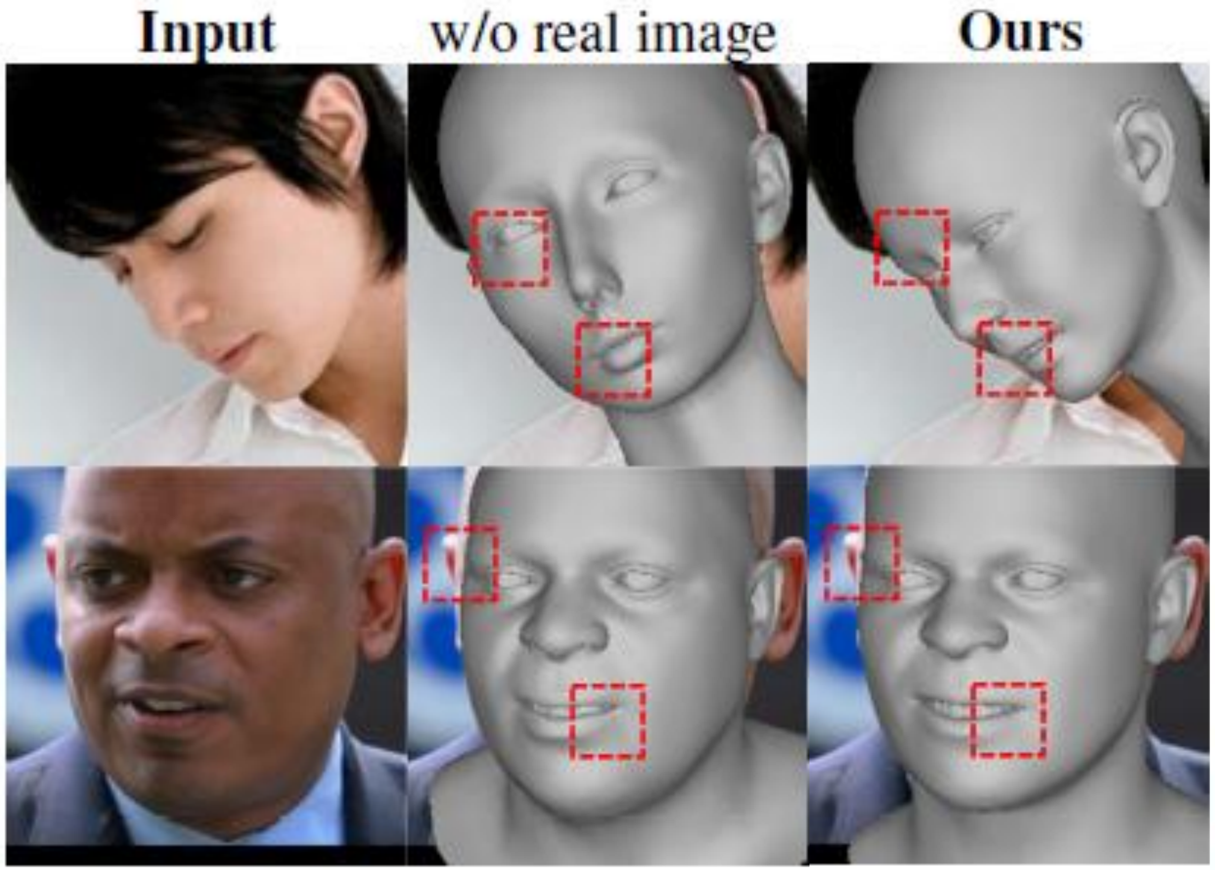


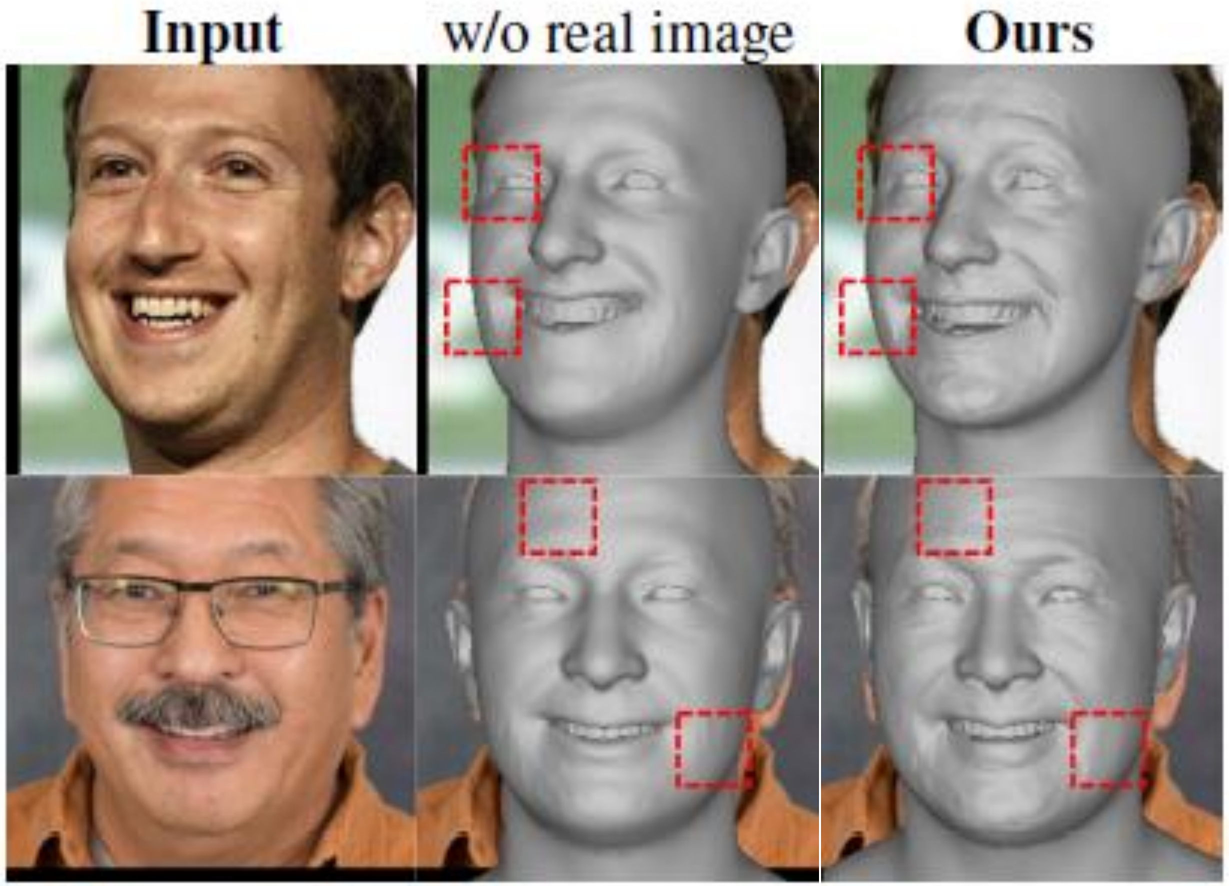
5. Ablation studies

Ablation Studies on loss functions and datasets

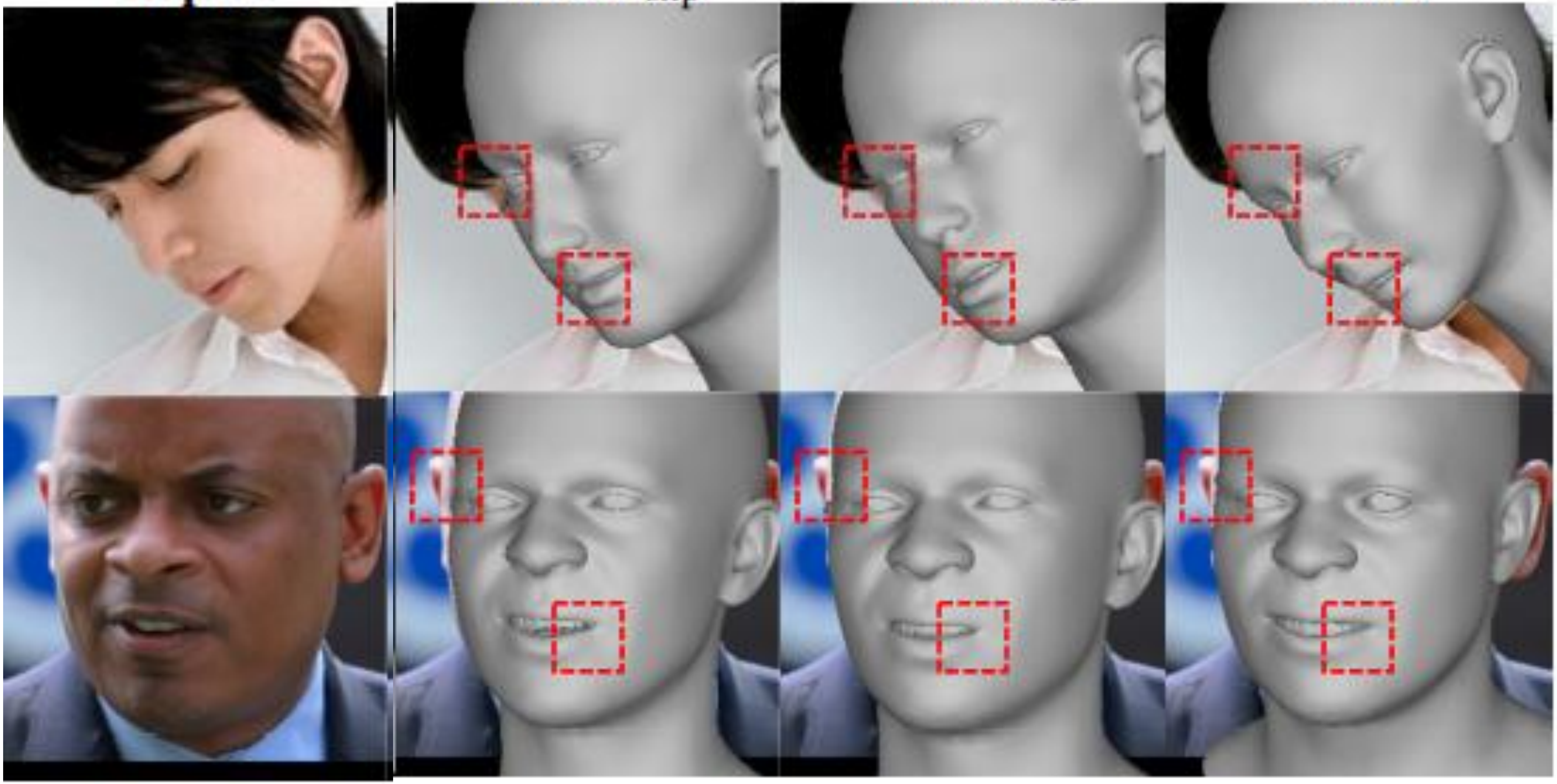
- Importance of real-world dataset
  - trained by synthetic dataset only
  - synthetic dataset + real-world dataset (self supervision)
- Importance of Loss functions
  - Coarse shape: w/o L_{shp}, w/o L_{kl} (overfitting prevention in L_{shp})
  - Detail reconstruction: w/o L_{detail}, w/o L_{kd}(knowledge distillation to static detail coefficient)
- Qualitative Experiment

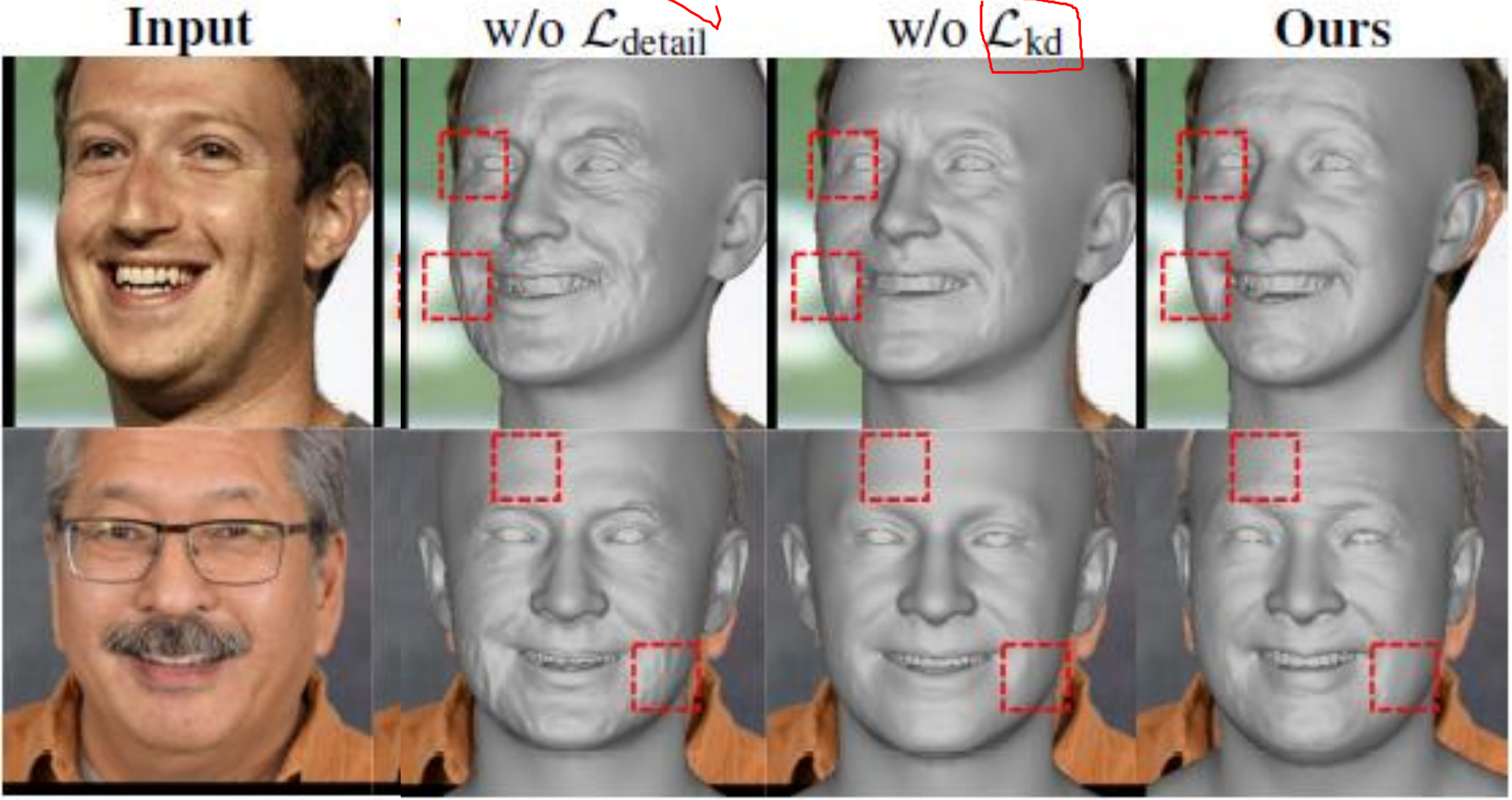
Ablation Studies on loss functions and datasets-Result





Input	w/o $\mathcal{L}_{shp}$	w/o $\mathcal{L}_{kl}$	Ours

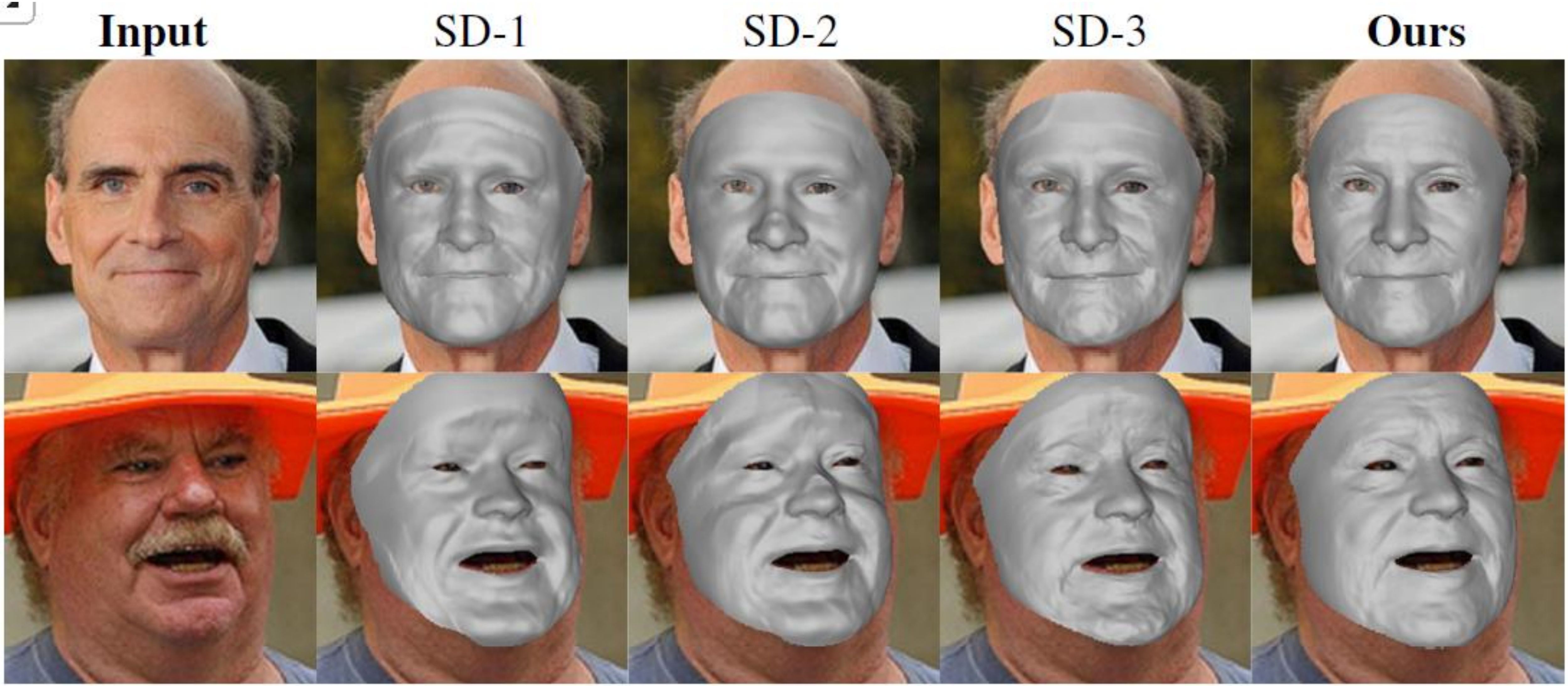



Input
w/o $\mathcal{L}_{\text{detail}}$
w/o $\mathcal{L}_{\text{kd}}$
Ours

Ablation Studies on SD-DeTail

- SD-1: directly generate D_{dyn} (learn without interpolating)
- SD-2: directly generate D_{com} and D_{str} , use interpolation
- SD-3: directly generate D_{sta}
- Ours: SD-DeTail

Ablation Studies on SD-DeTail -result



SD-1: directly generate D_{dyn} (learn without interpolating)
SD-2: directly generate D_{com} and D_{str} , use interpolation
SD-3: directly generate D_{sta}
Ours: SD-DeTail

6.Conclusion

Summary

- HiFace: 3D face regeneration from single image to high-fidelity 3D face
- based on 3DMMs.
- simplify fine detail generation problem as regression and interpolation tasks
- SD-DeTail module: decouple static and dynamic detail
- vertex tension: interpolates dynamic detail from expression
- hybrid dataset: synthetic GT + real-world self-supervision
- new loss function: learn coarse shape & fine detail simultaneously