JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

SOUTHEAST UNIVERSITY
東南大學

Semi-supervised Hand Appearance Recovery via Structure Disentanglement and Dual Adversarial Discrimination

Zimeng Zhao, Binghui Zuo, Zhiyu Long and Yangang Wang

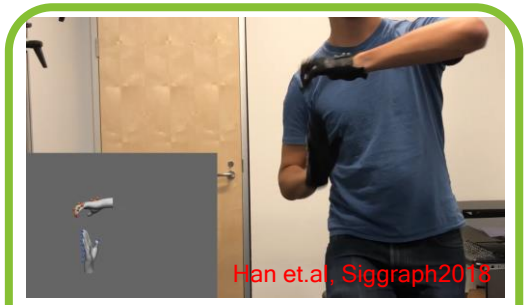
Southeast University

Project Page

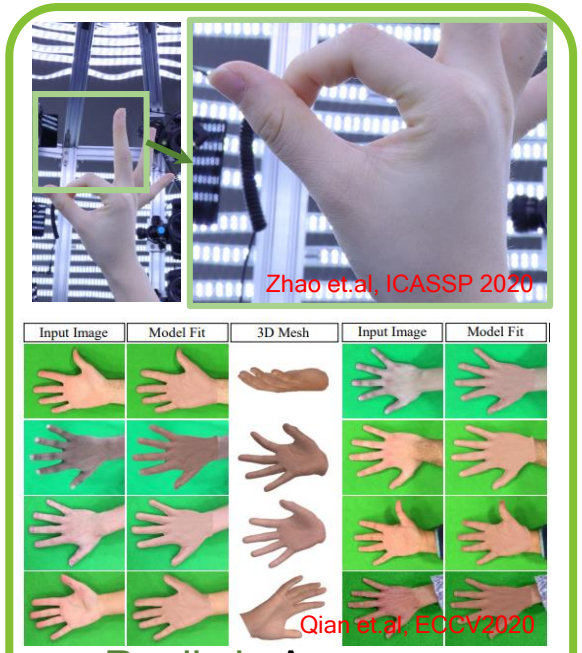
WED-AM-372

Human Hand Data Capture

Separate Acquisition

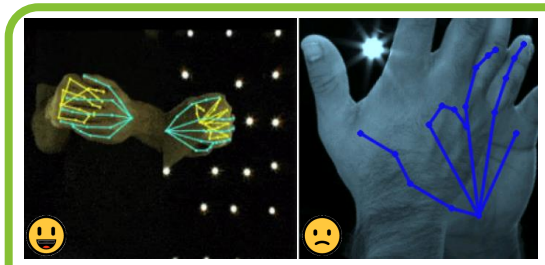


Accurate Pose:
Marker-based Capture

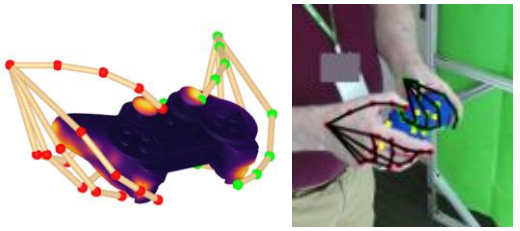


Realistic Appearance:
Bare Skin Capture

Synchronous Acquisition

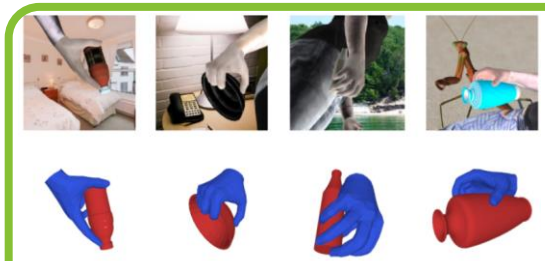


Moon et.al, ECCV 2020



Brahmbhatt et.al, ECCV 2020

Marker-less Procedure with Pose Estimator



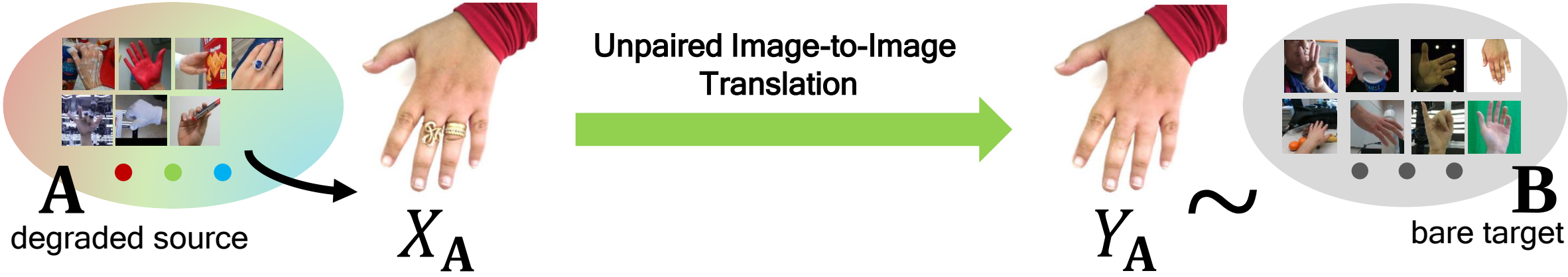
Hasson et.al, CVPR 2022

Gao et.al, NeurIPS 2022

Synthetic Procedure with Differentiable Rendering

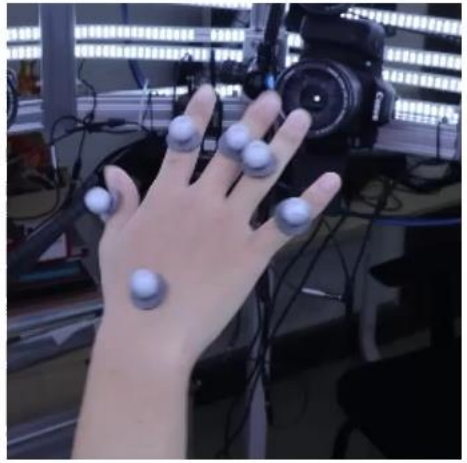
- Pose & Appearance **Synchronous** Acquisition
  - Marker-less Procedure: Depends on the estimator stability
  - Synthetic Procedure: Suffers from the rendering realness

Insight

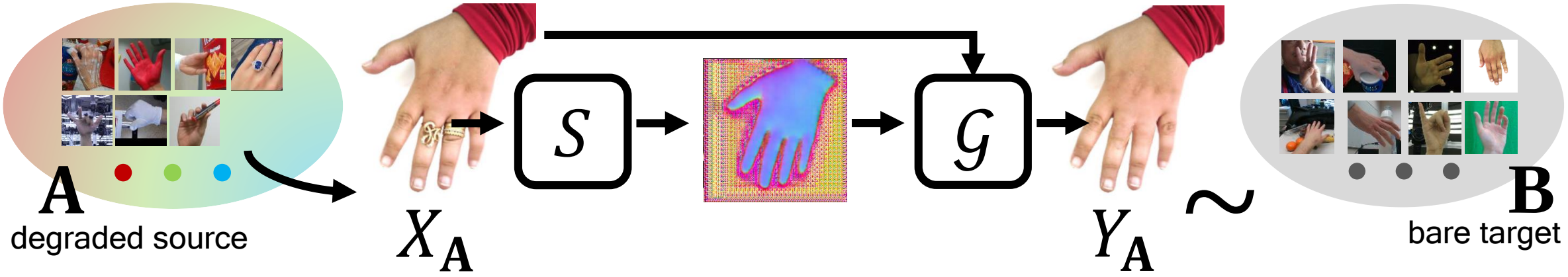

A
degraded source
X_A
Unpaired Image-to-Image Translation
Y_A
~
B
bare target



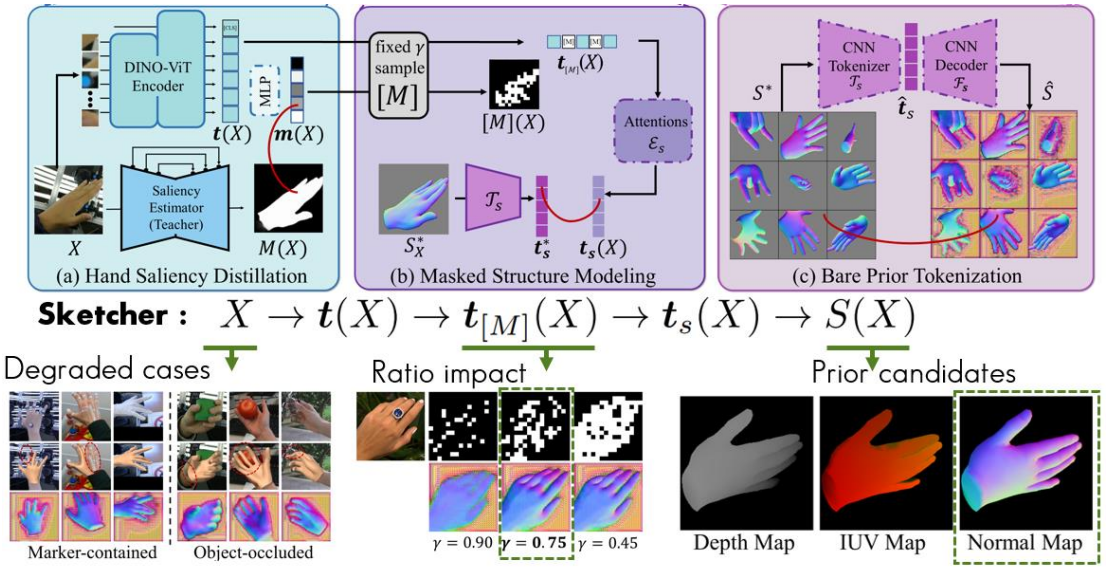




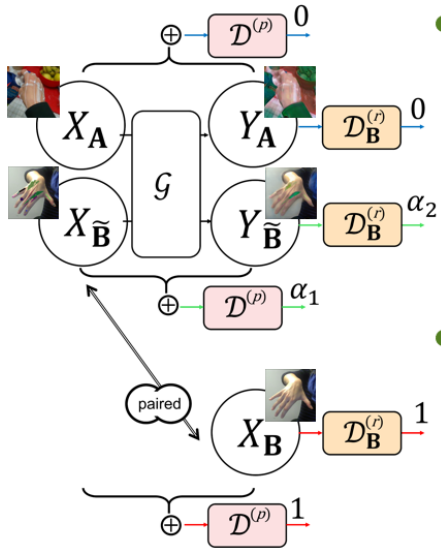
Overview



Structure Disentanglement



Dual Adversarial Discrimination



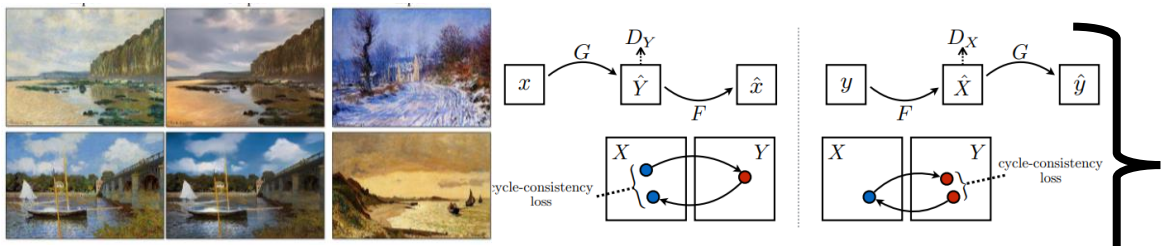
- **Translation loss**

$$
\begin{aligned}
L_{\mathcal{G}} = & \|(Y_{\mathbf{A}}\text{-}X_A) \odot (1\text{-}M[S(X_{\mathbf{A}})])\|_F \\
& + \|(Y_{\tilde{\mathbf{B}}}\text{-}X_{\tilde{\mathbf{B}}}) \odot (1\text{-}M[S(X_{\tilde{\mathbf{B}}})])\|_F \\
& + |\mathcal{D}_{\mathbf{B}}^{(r)}(Y_{\mathbf{A}})\text{-}1| + |\mathcal{D}_{\mathbf{B}}^{(r)}(Y_{\tilde{\mathbf{B}}})\text{-}1| \\
& + |\mathcal{D}^{(p)}(X_{\mathbf{A}} \oplus Y_{\mathbf{A}})\text{-}1| + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus Y_{\tilde{\mathbf{B}}})\text{-}1|
\end{aligned}
$$

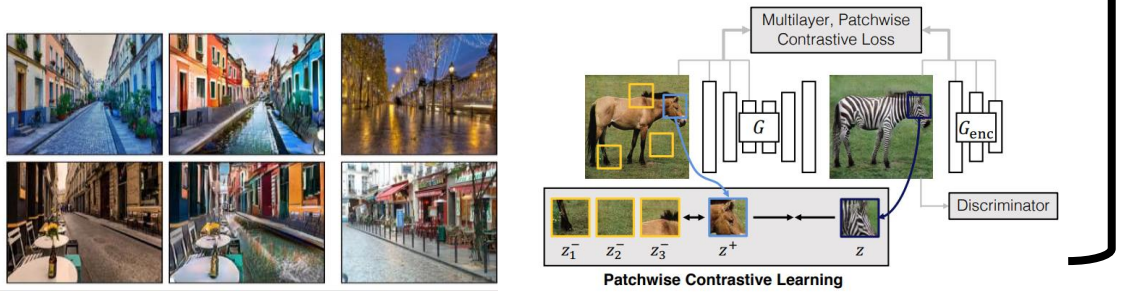
- **Dual-discrimination loss**

$$
\begin{cases}
L_{\mathcal{D}}^{(r)} = |\mathcal{D}_{\mathbf{B}}^{(r)}(\underline{Y_{\mathbf{A}}})\text{-}0| + |\mathcal{D}_{\mathbf{B}}^{(r)}(\underline{Y_{\tilde{\mathbf{B}}}})\text{-}\alpha_2| + |\mathcal{D}_{\mathbf{B}}^{(r)}(X_{\mathbf{B}})\text{-}1| \\
L_{\mathcal{D}}^{(p)} = |\mathcal{D}^{(p)}(X_{\mathbf{A}} \oplus \underline{Y_{\mathbf{A}}})\text{-}0| + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus \underline{Y_{\tilde{\mathbf{B}}}})\text{-}\alpha_1| \\
\qquad + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus X_{\mathbf{B}})\text{-}1|
\end{cases}
$$

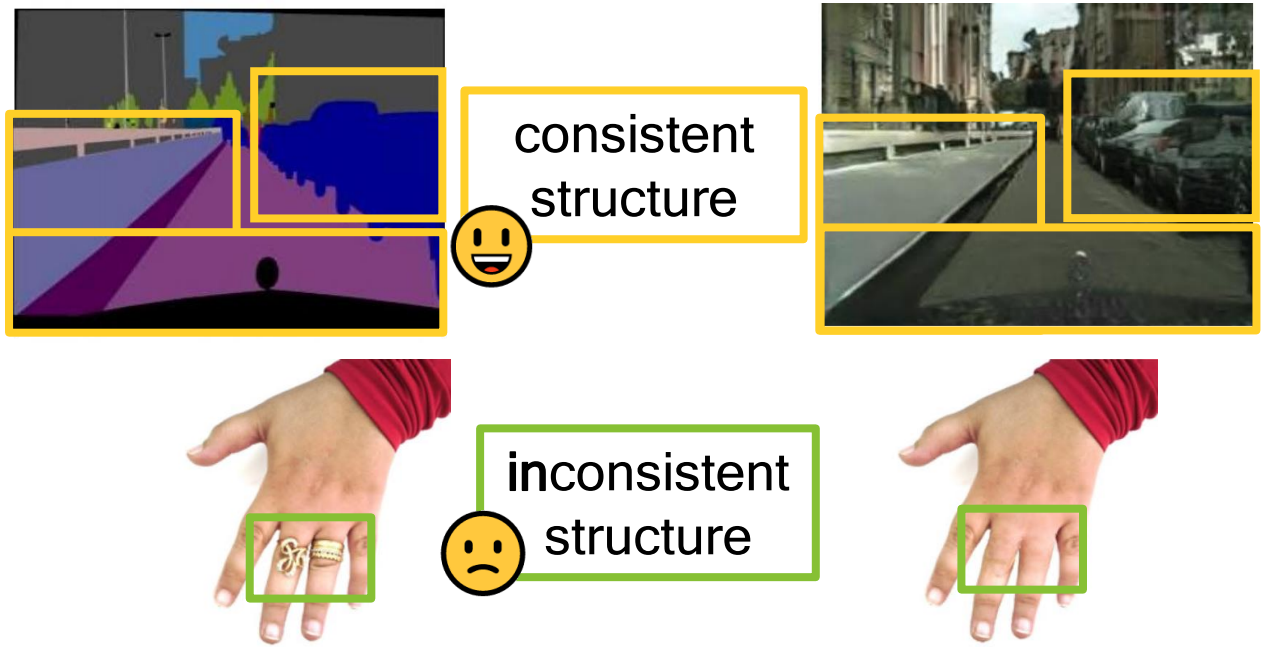
Challenges



CycleGAN, Zhu et.al. ICCV 2017

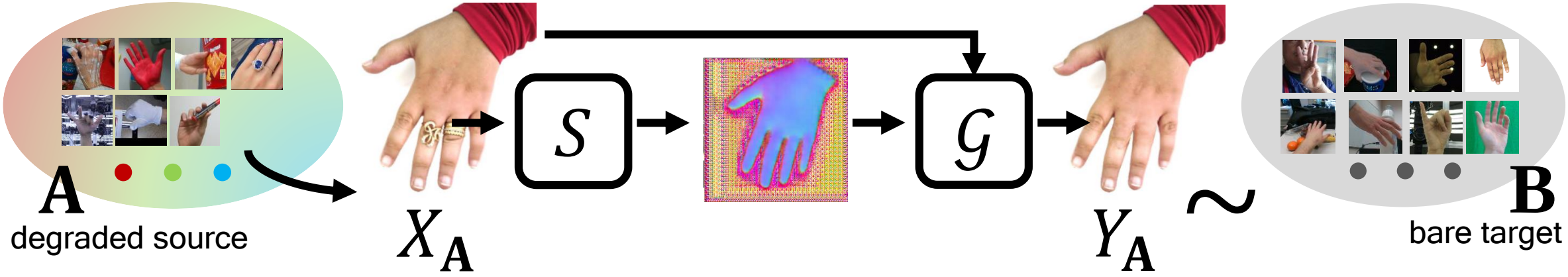


CUT, Park et.al. ECCV 2020



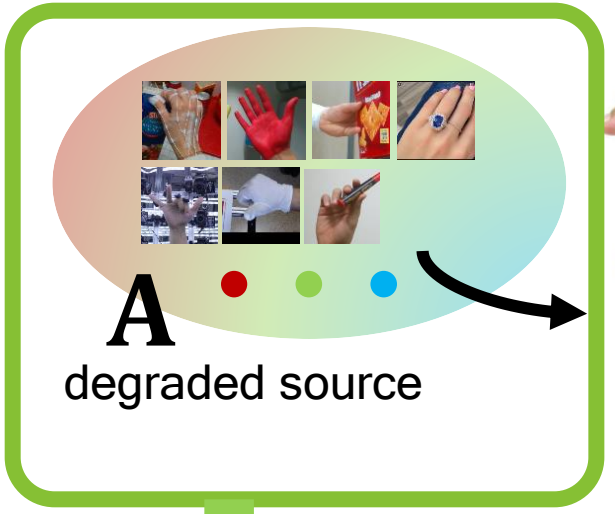
- **Data** : Degraded-bare hand image pairs are almost impossible to obtain.
- **Problem** : Background, foreground hand, and foreground markers require different processing options.
- **Baseline** : Existing unsupervised translators cannot handle the problem with partial structure inconsistency.

Contributions



- A semi-supervised framework that makes degraded images in marker-based MoCap regain bare appearance.
- A powerful ViT sketcher that disentangles bare hand structure without parametric model dependencies.
- An adversarial scheme that promotes the degraded-to-bare appearance wrapping effectively.

Step 1. Structure Disentanglement

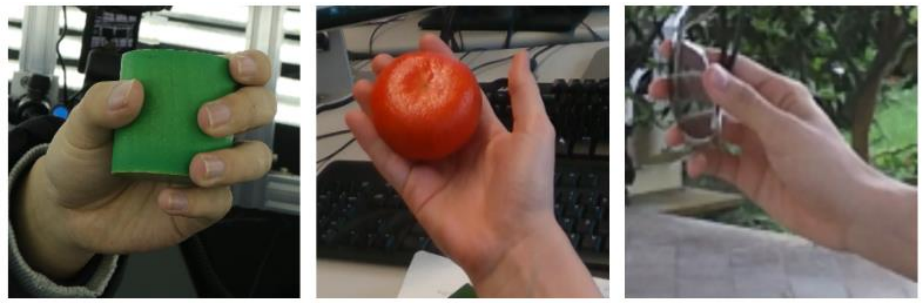


X_A

Y_A ~

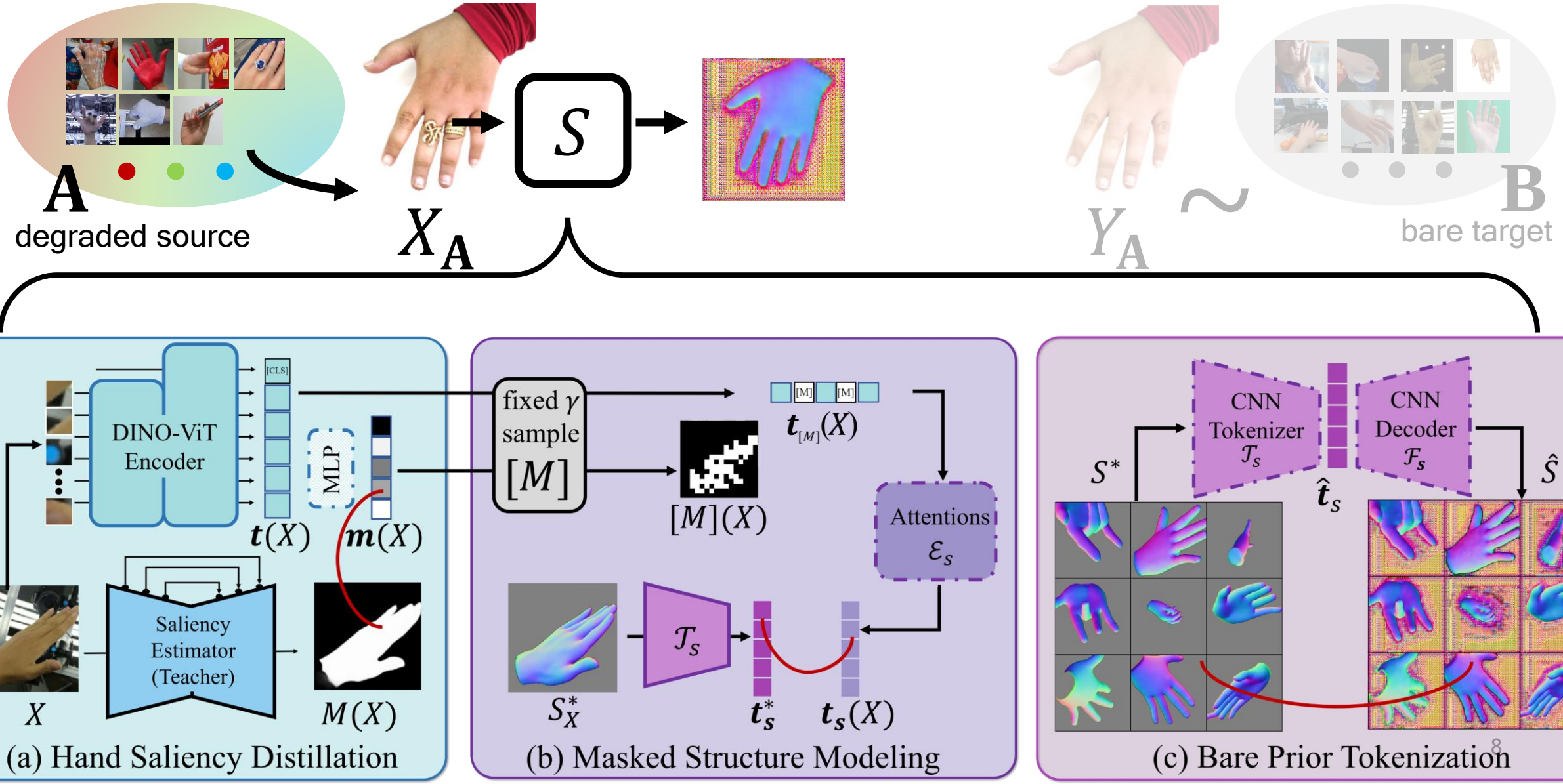
Degraded Cases We Concerned :

Marker-contained

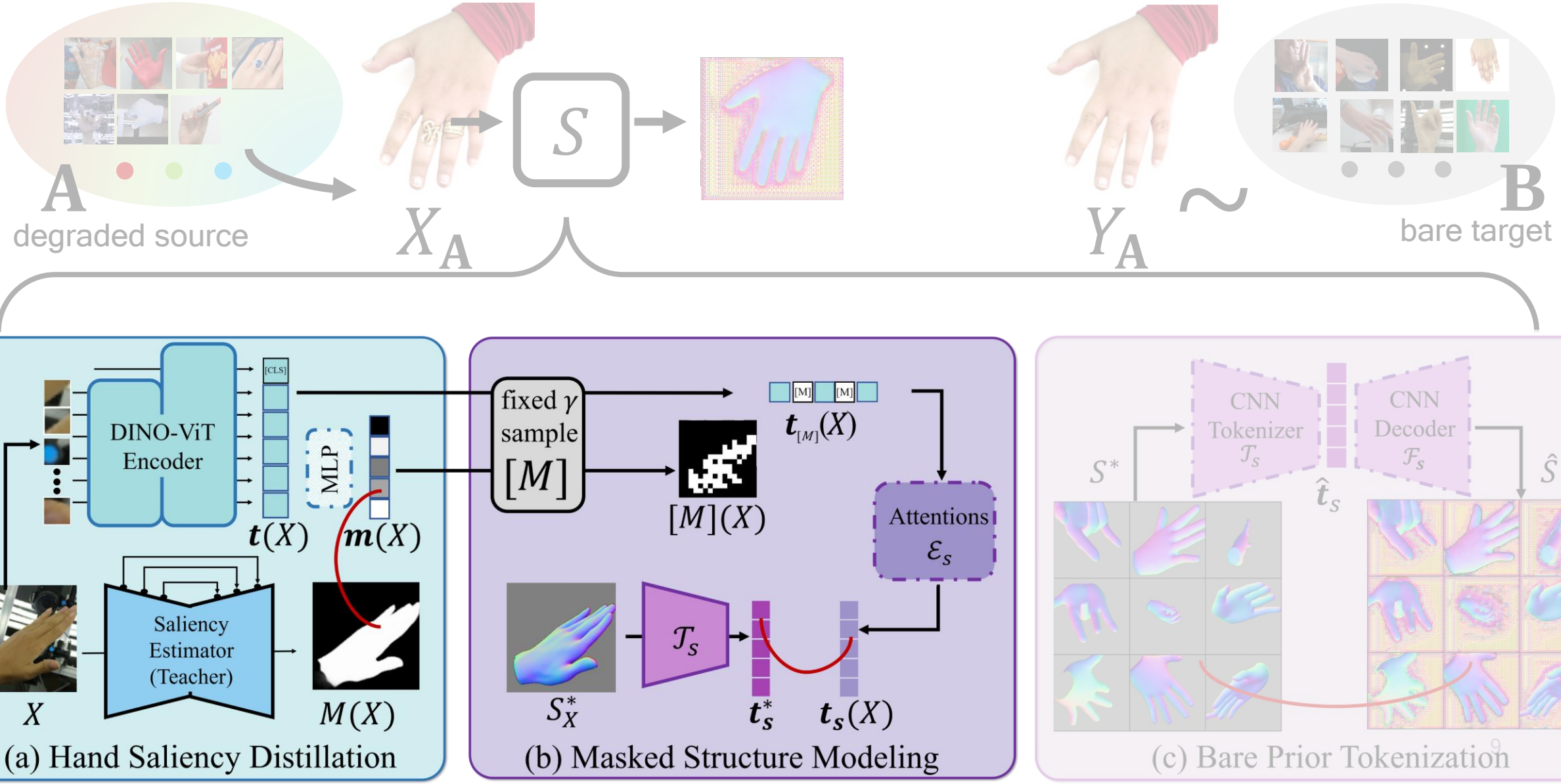


Object-occluded

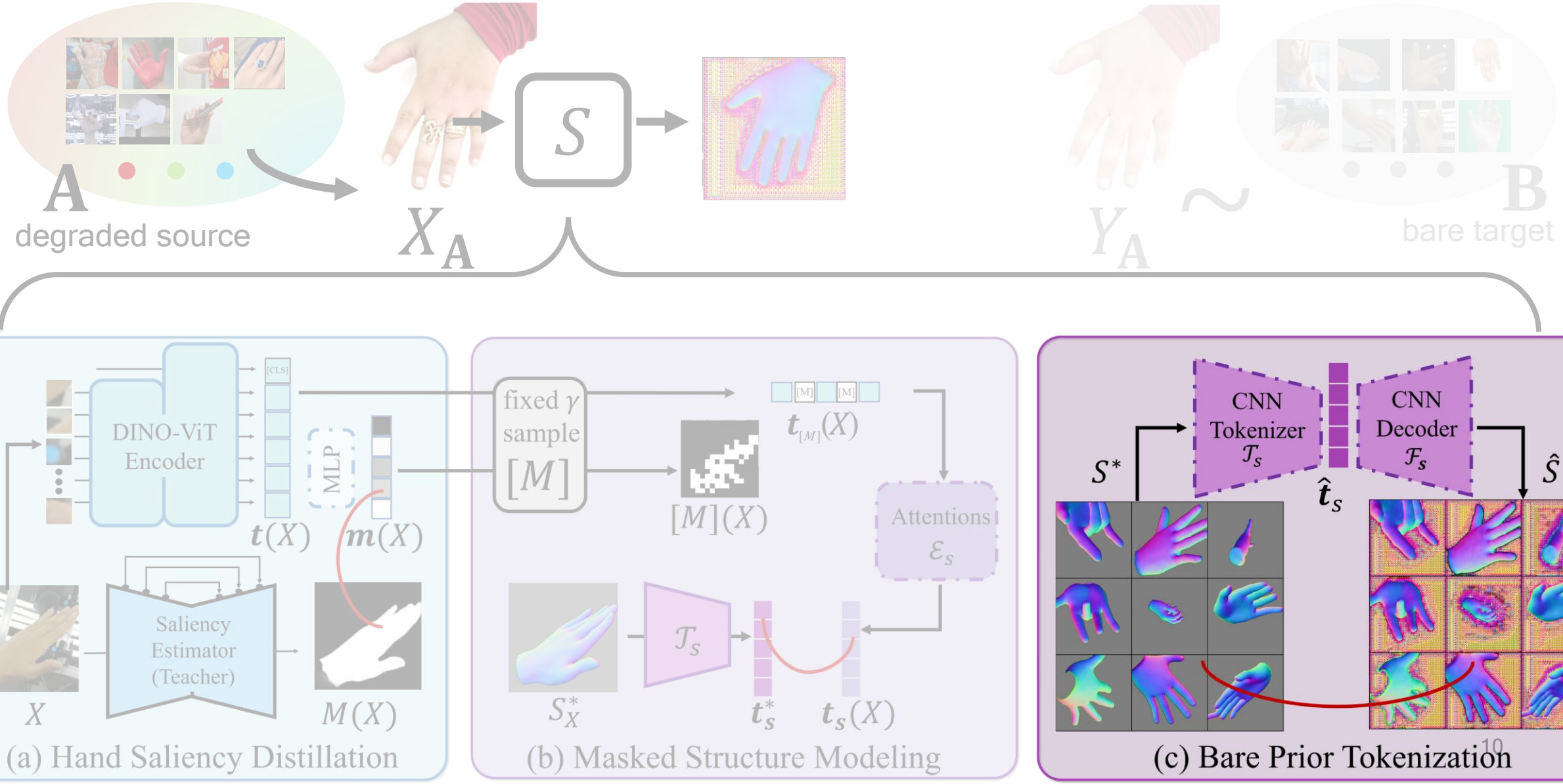
Step 1. Structure Disentanglement



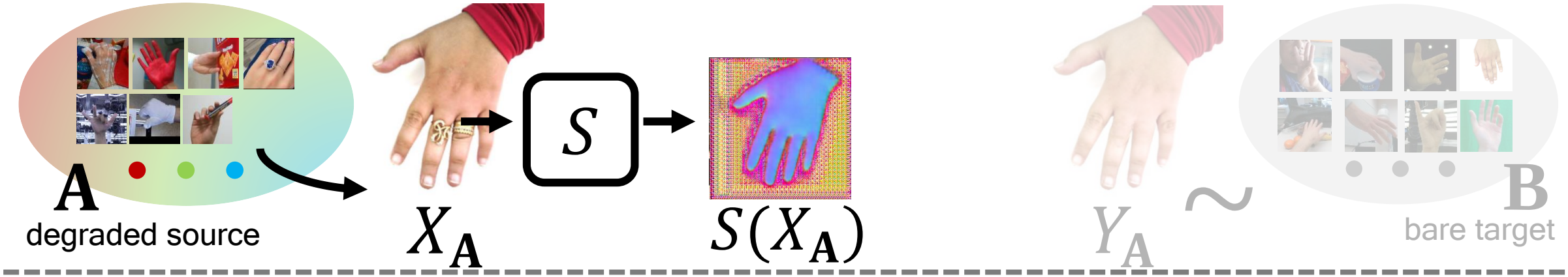
Step 1. Structure Disentanglement



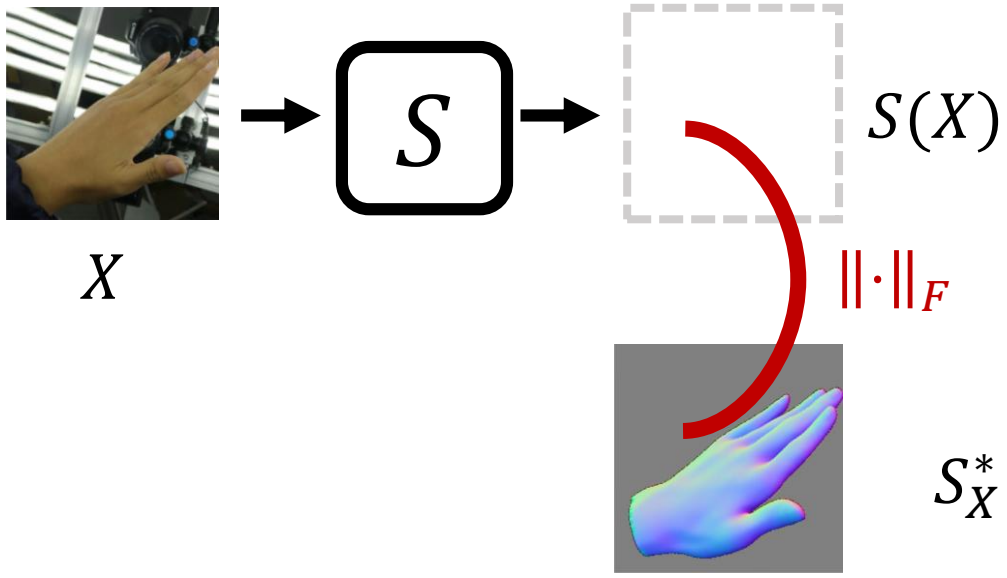
Step 1. Structure Disentanglement



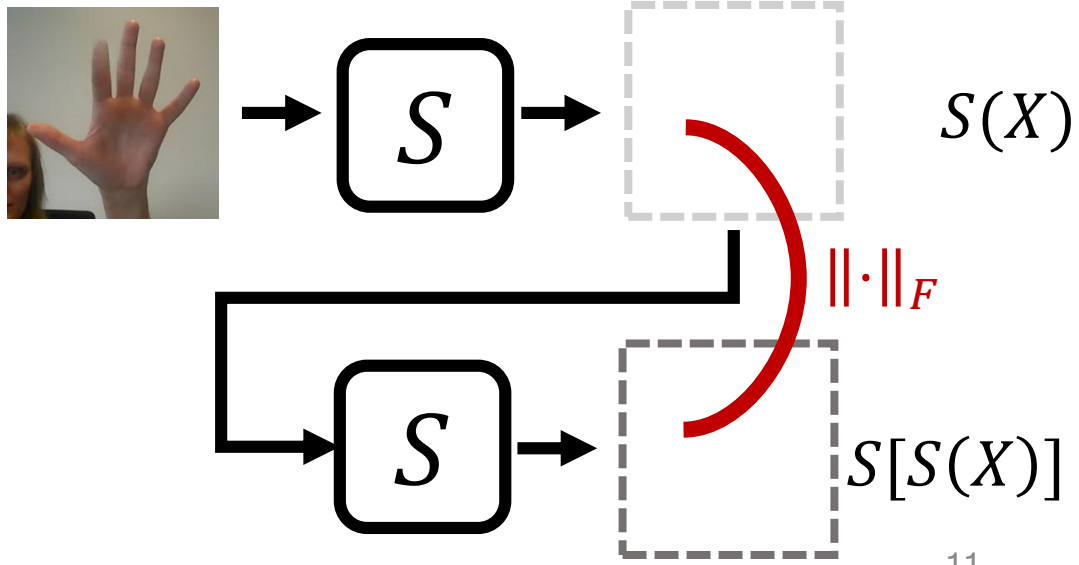
Step 1. Structure Disentanglement



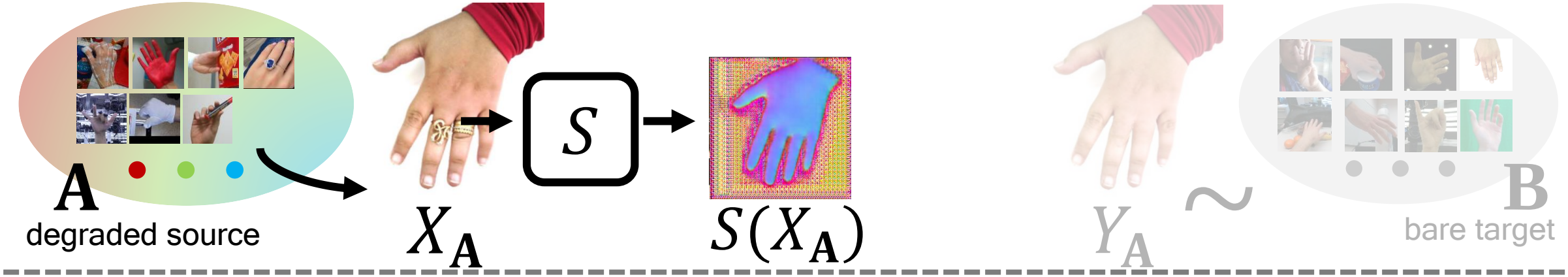
Train with **labeled** data $\{X, S_X^*\}$

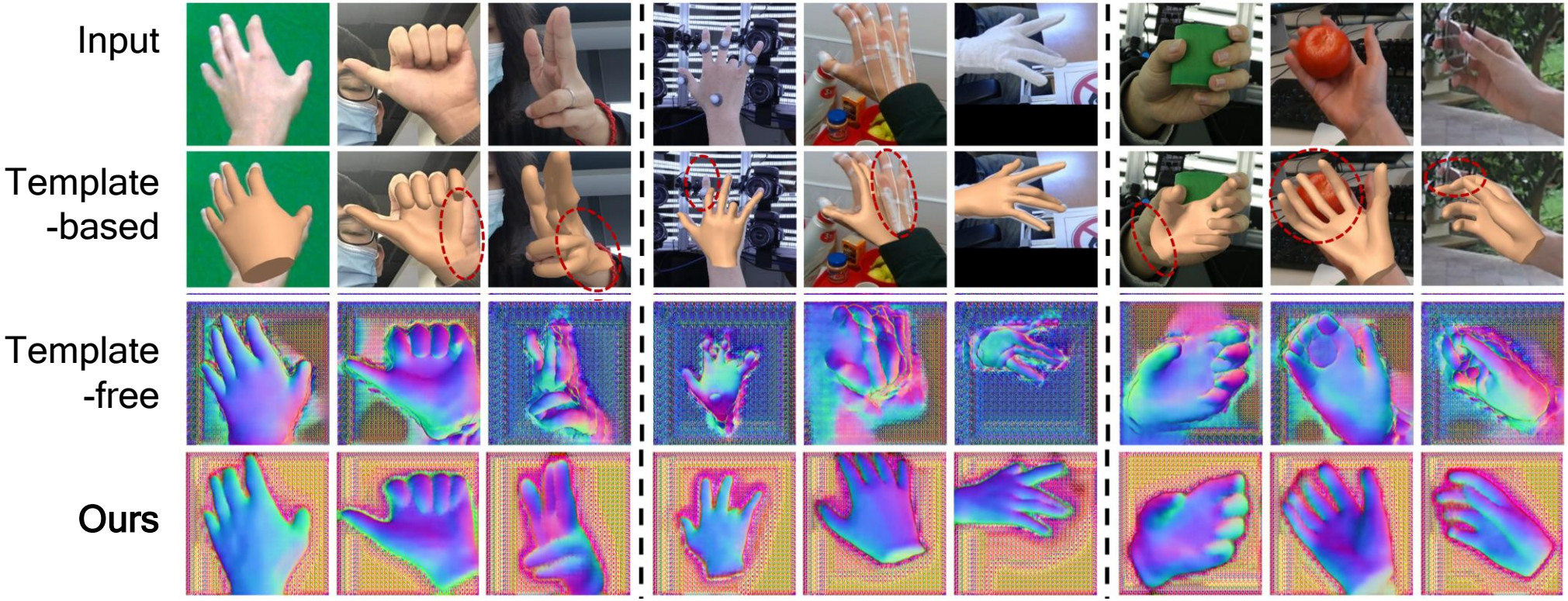


Train with **unlabeled** data $\{X\}$

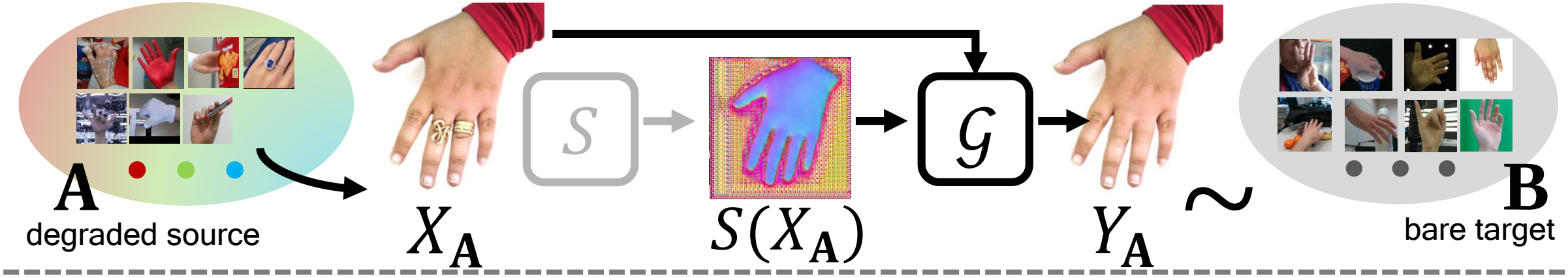


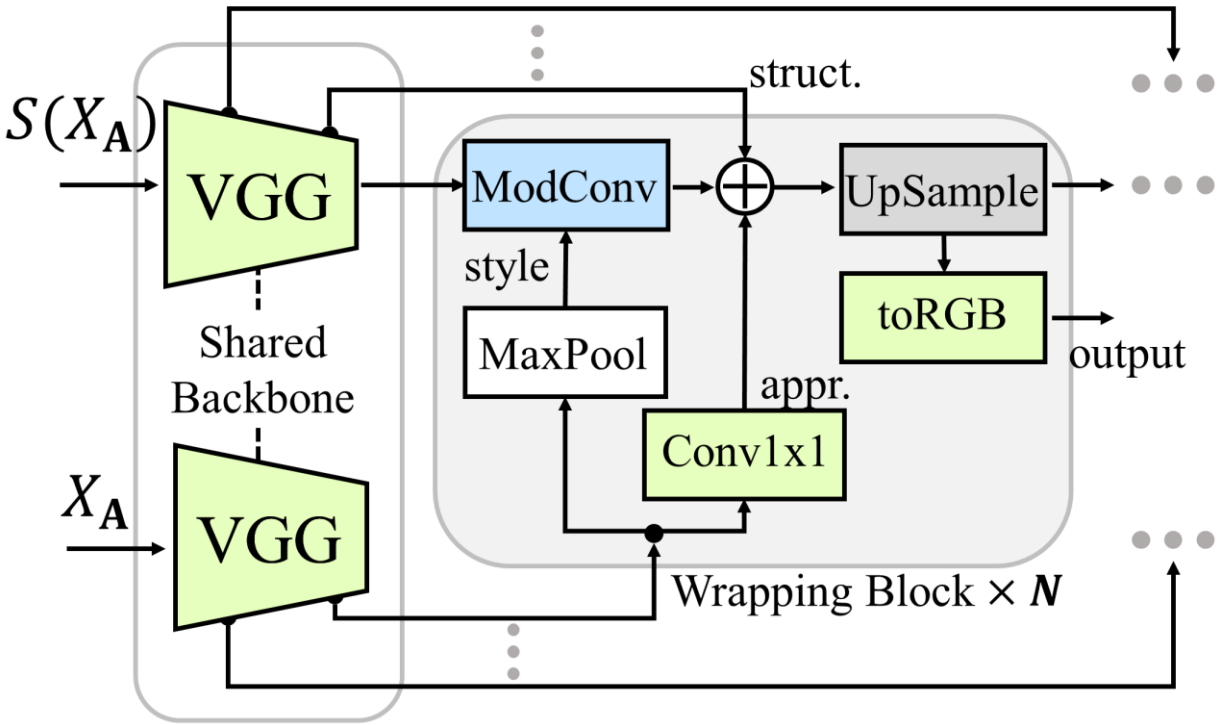
Step 1. Structure Disentanglement



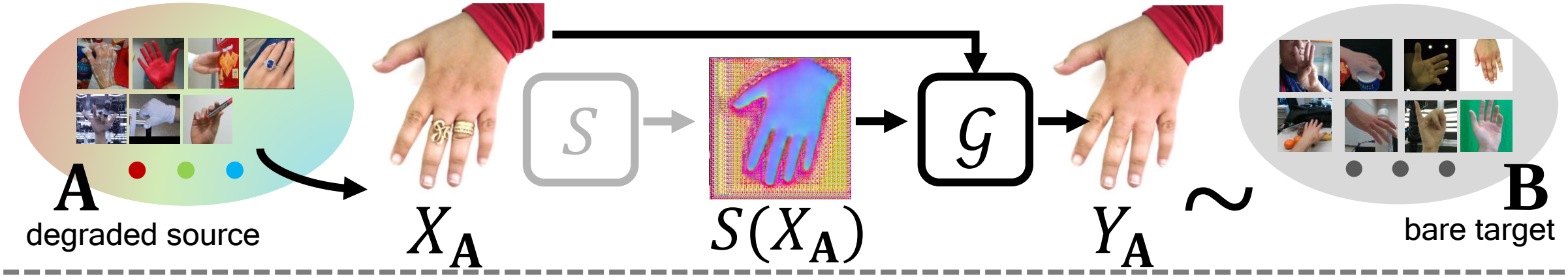


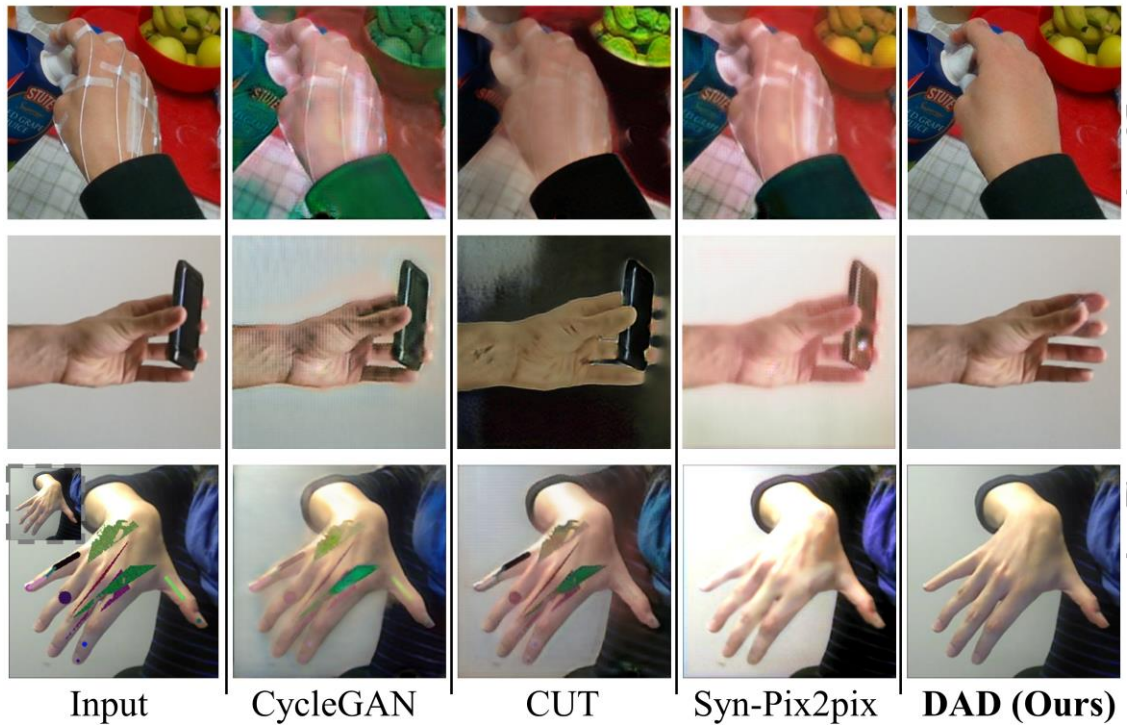
Step 2. Appearance Wrapping

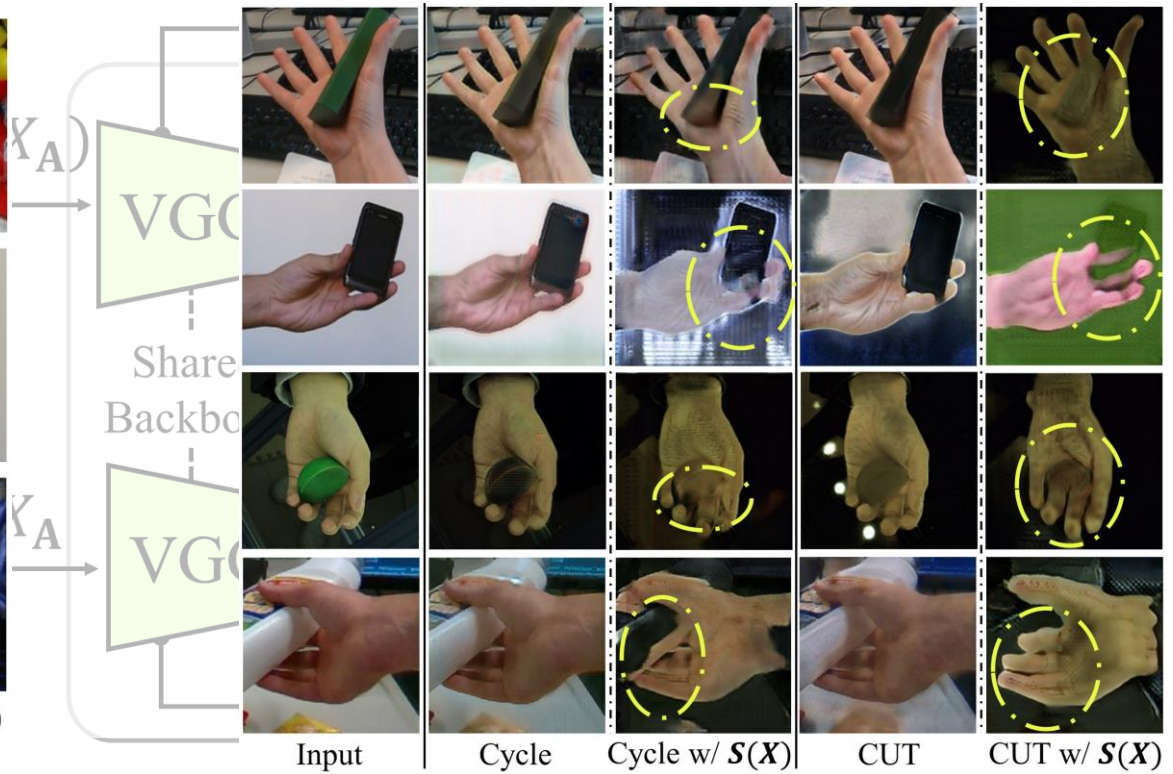




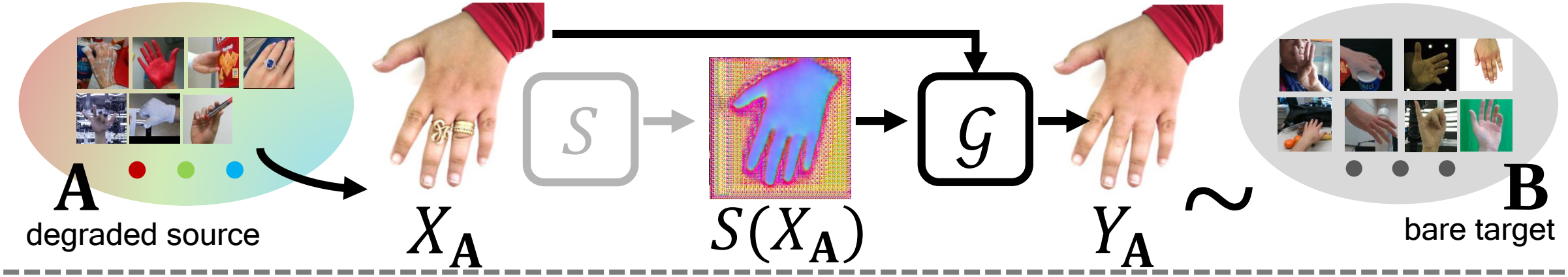
Step 2. Appearance Wrapping

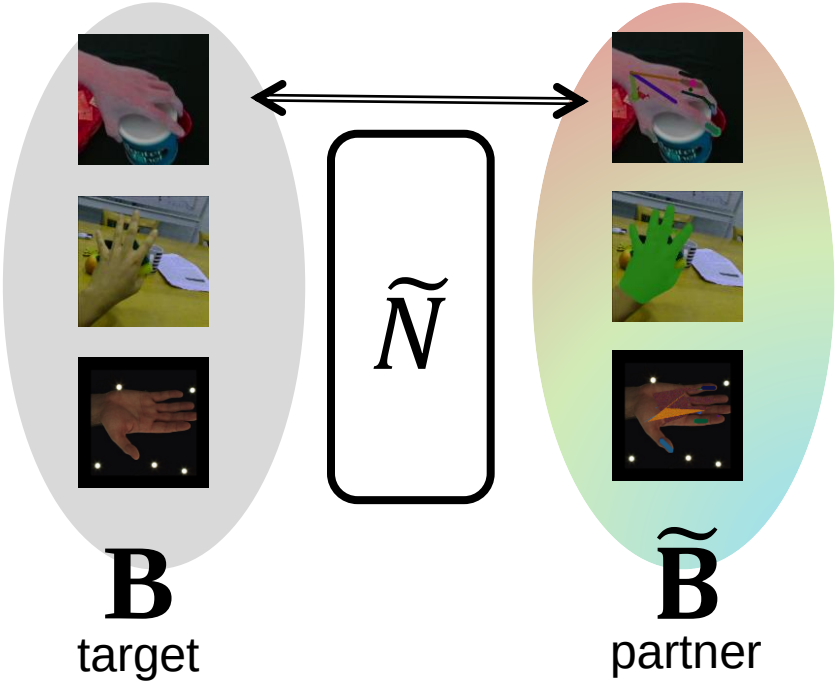


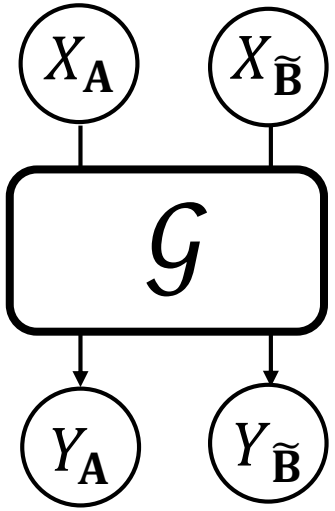




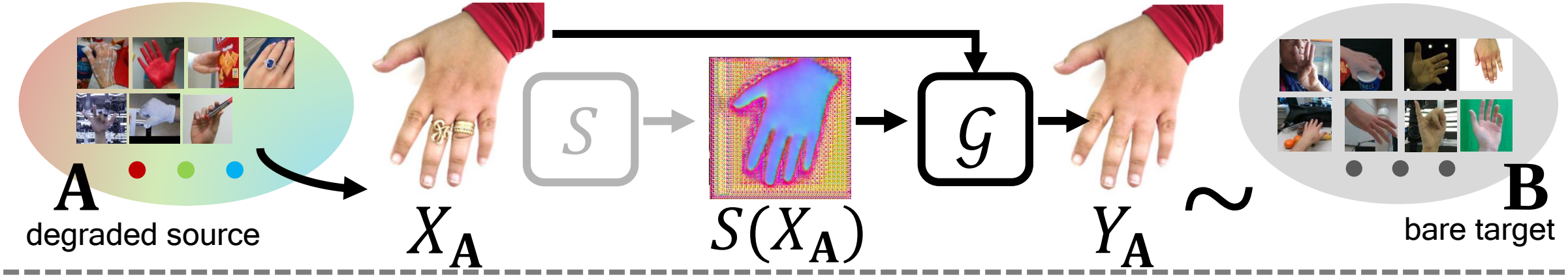
Step 2. Appearance Wrapping

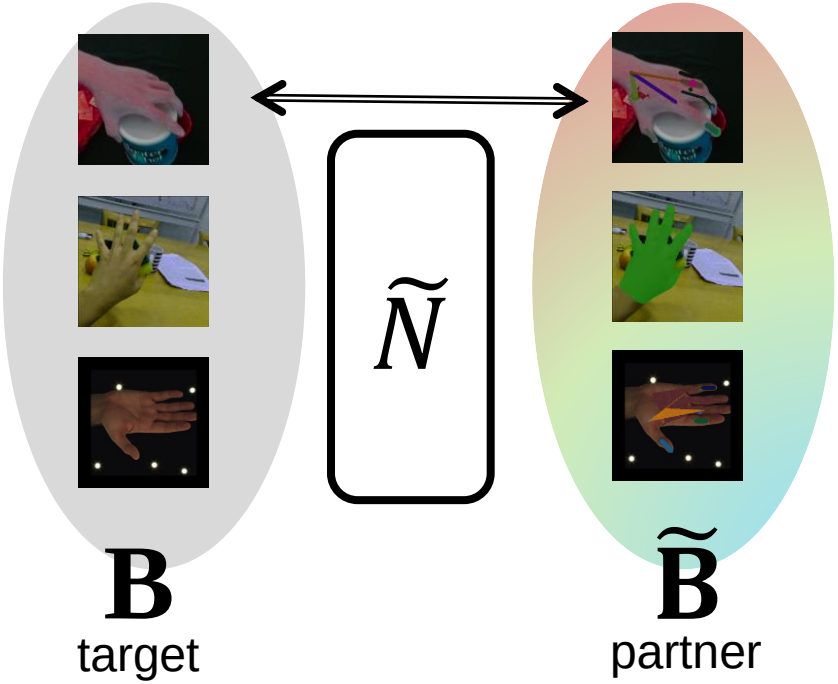


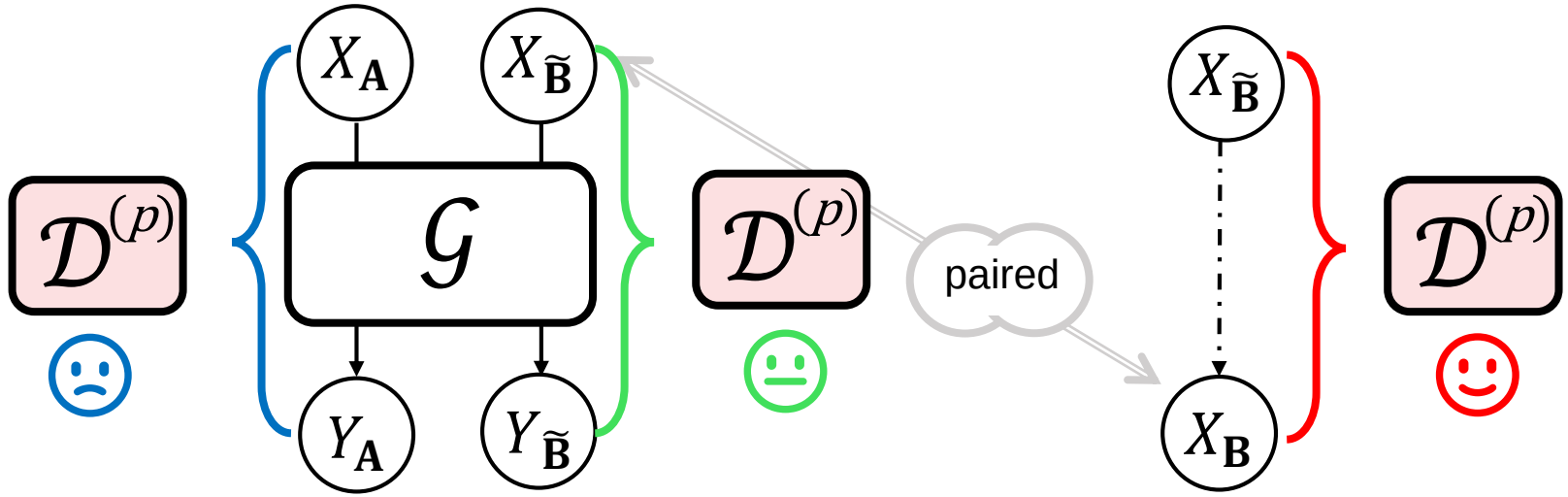




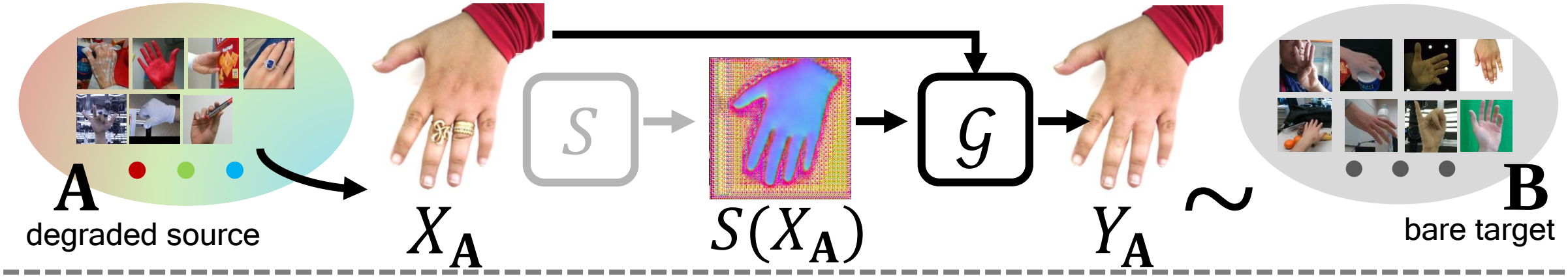
Step 2. Appearance Wrapping

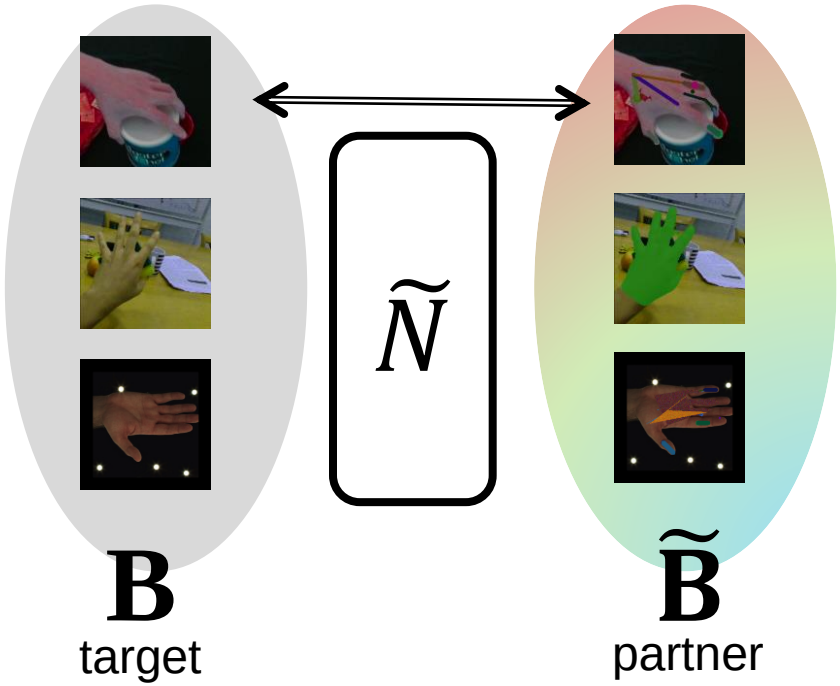


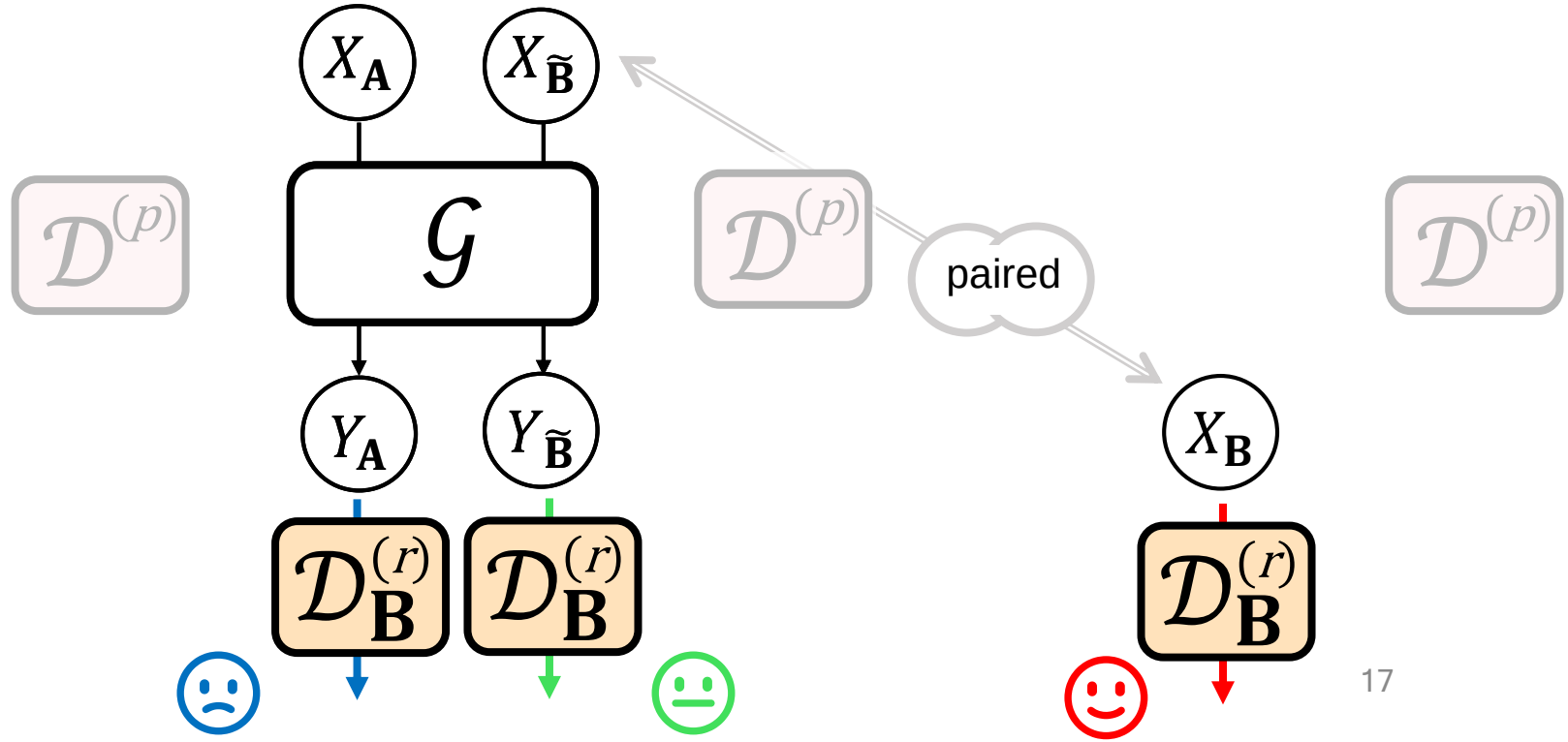




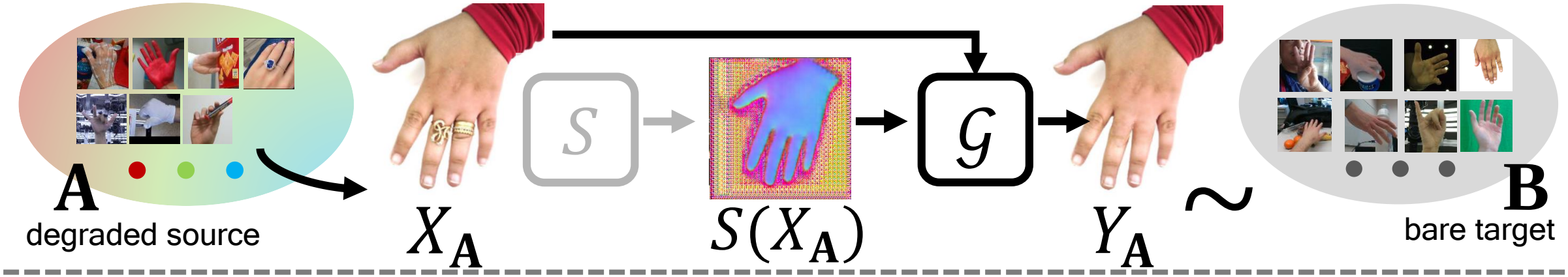
Step 2. Appearance Wrapping

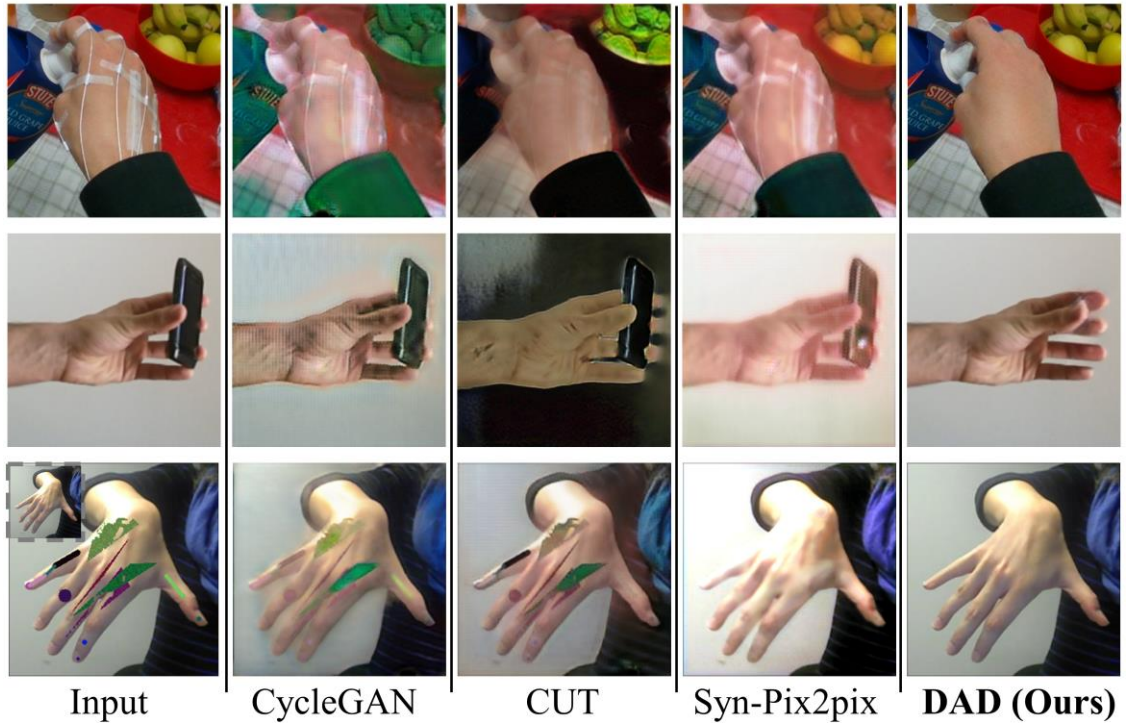


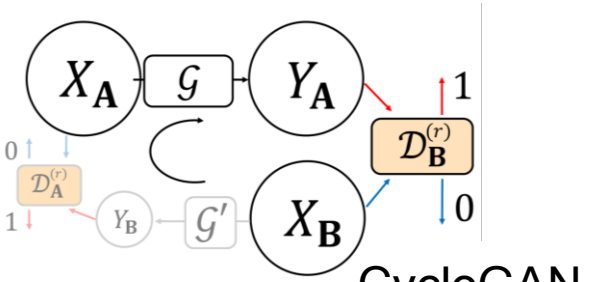




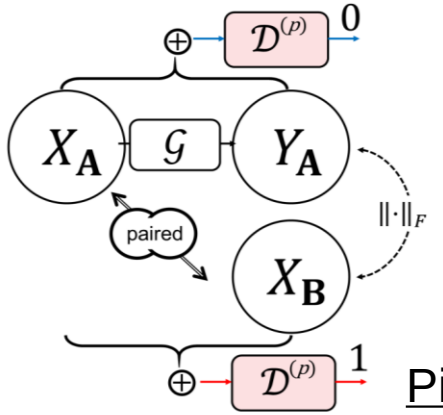
Step 2. Appearance Wrapping



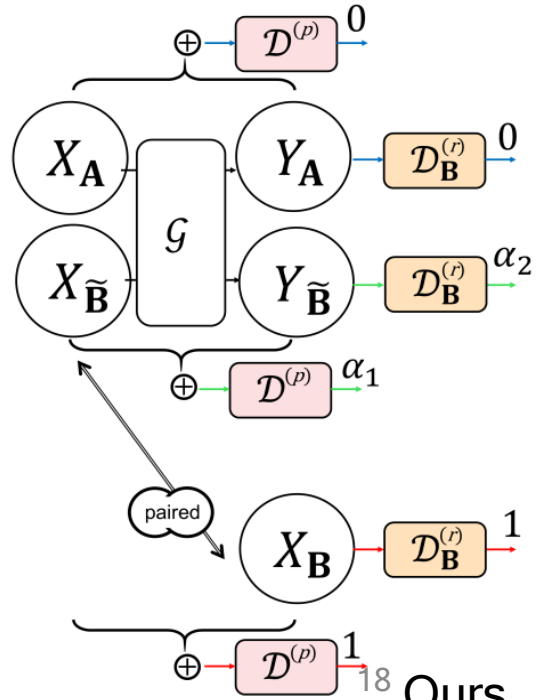




CycleGAN

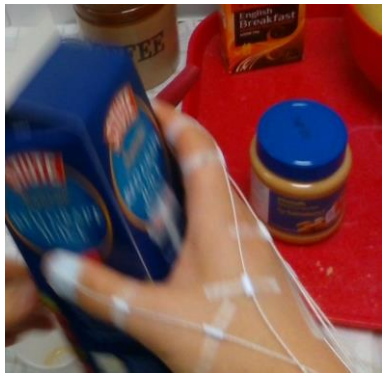


Pix2pix



Ours

Qualitative Results on Translation $\mathbf{A_1} \rightarrow \mathbf{B}$

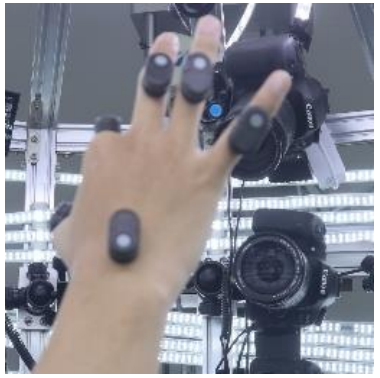





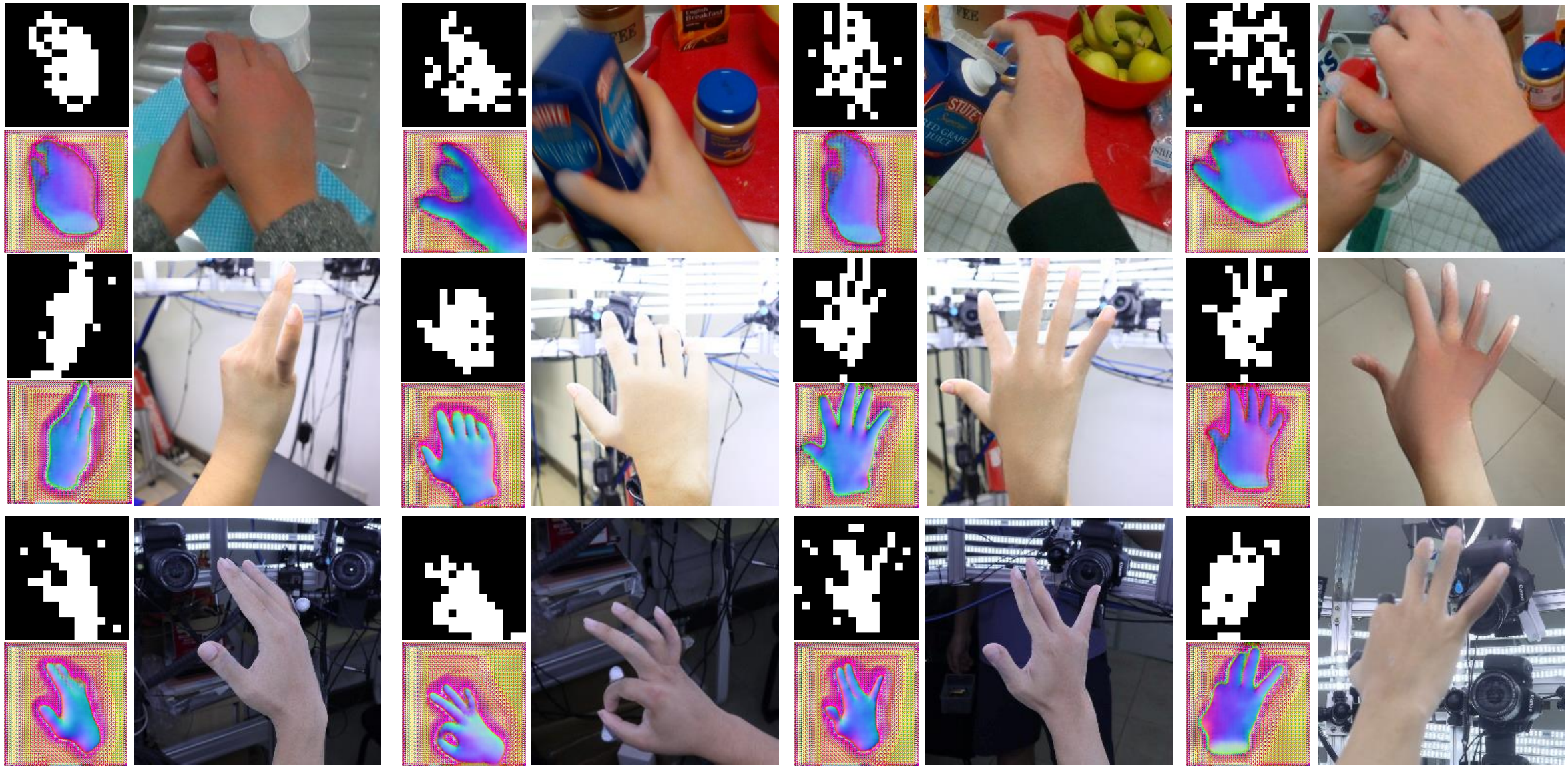






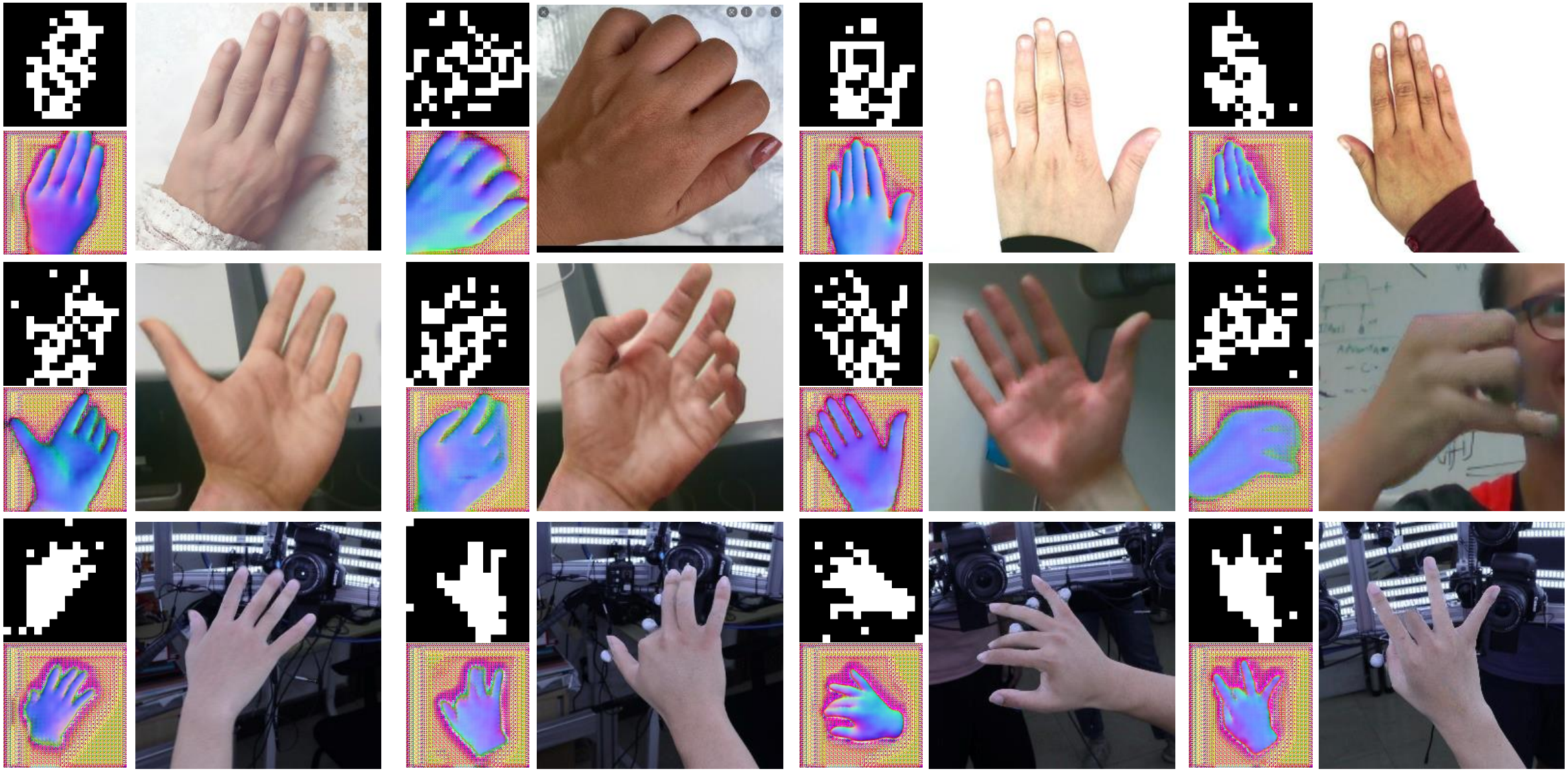


Qualitative Results on Translation $\mathbf{A_1} \rightarrow \mathbf{B}$



Qualitative Results on Translation $\mathbf{A_1} \rightarrow \mathbf{B}$

Qualitative Results on Translation $\mathbf{A_1 \rightarrow B}$

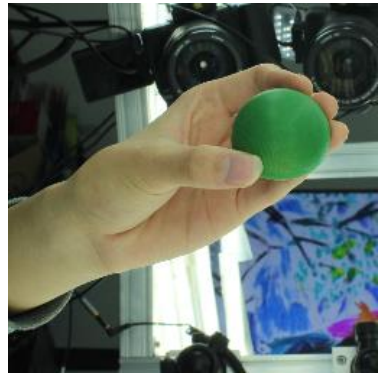


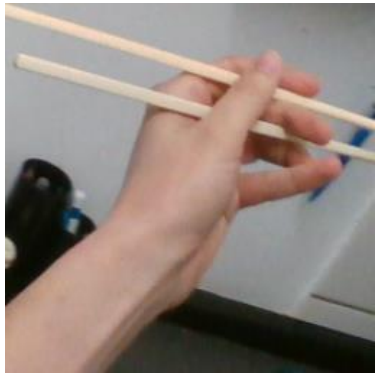
Qualitative Results on Translation $\mathbf{A_2 \rightarrow B}$

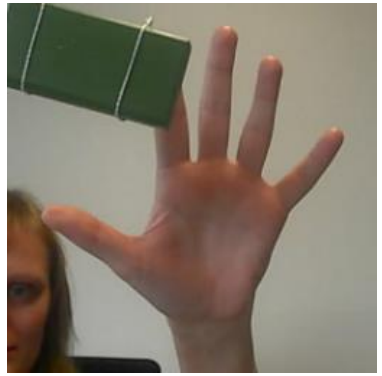




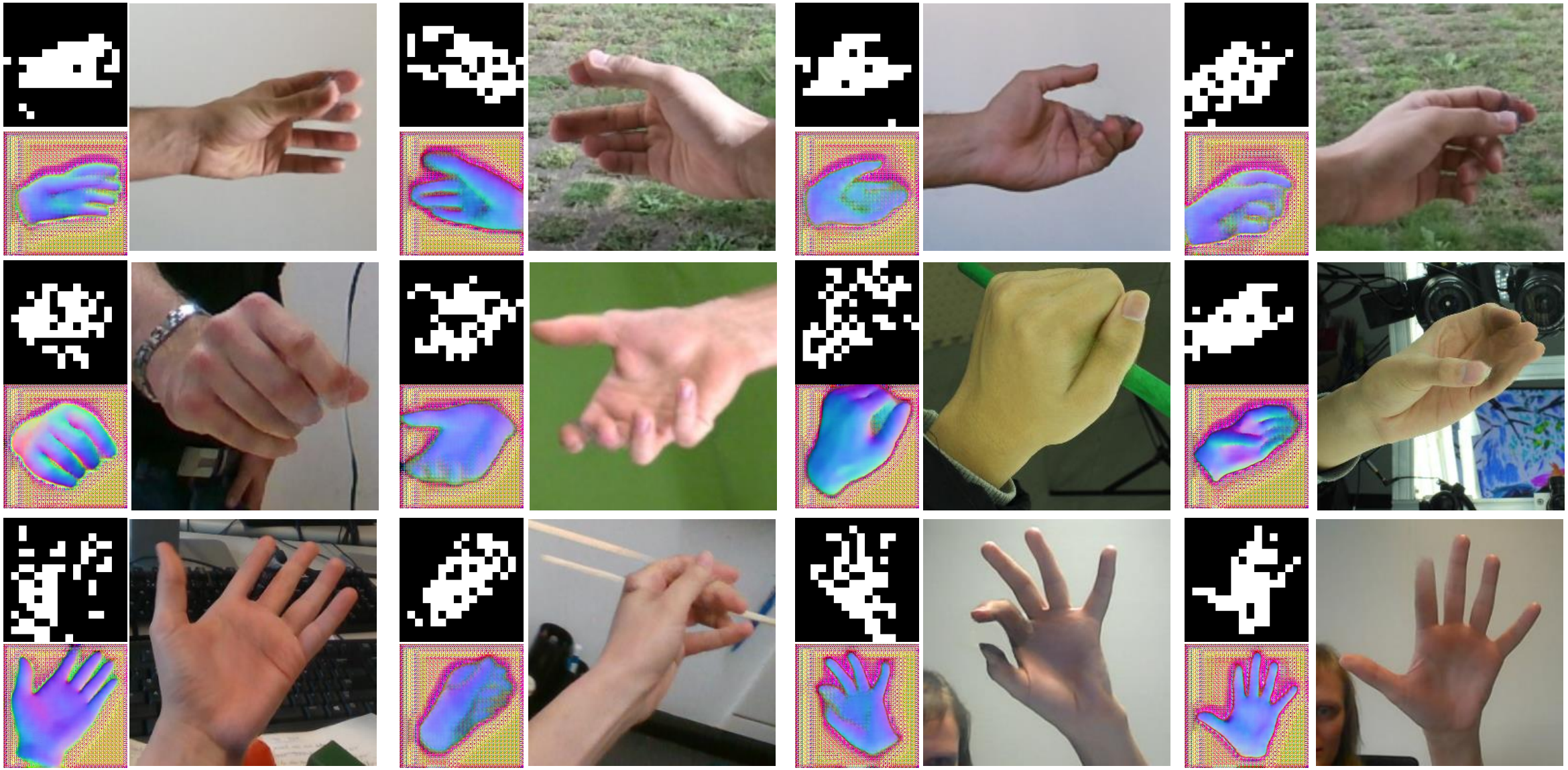






Qualitative Results on Translation $\mathbf{A_2 \rightarrow B}$

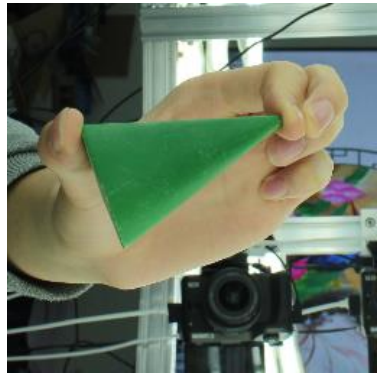


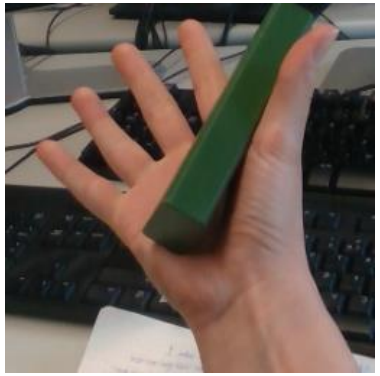
Qualitative Results on Translation $\mathbf{A_2 \rightarrow B}$

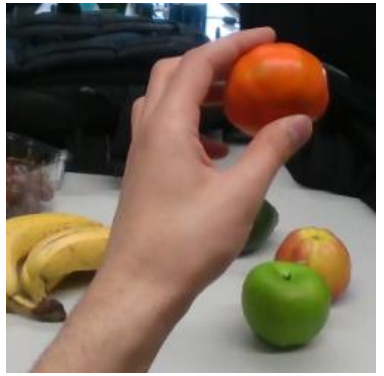




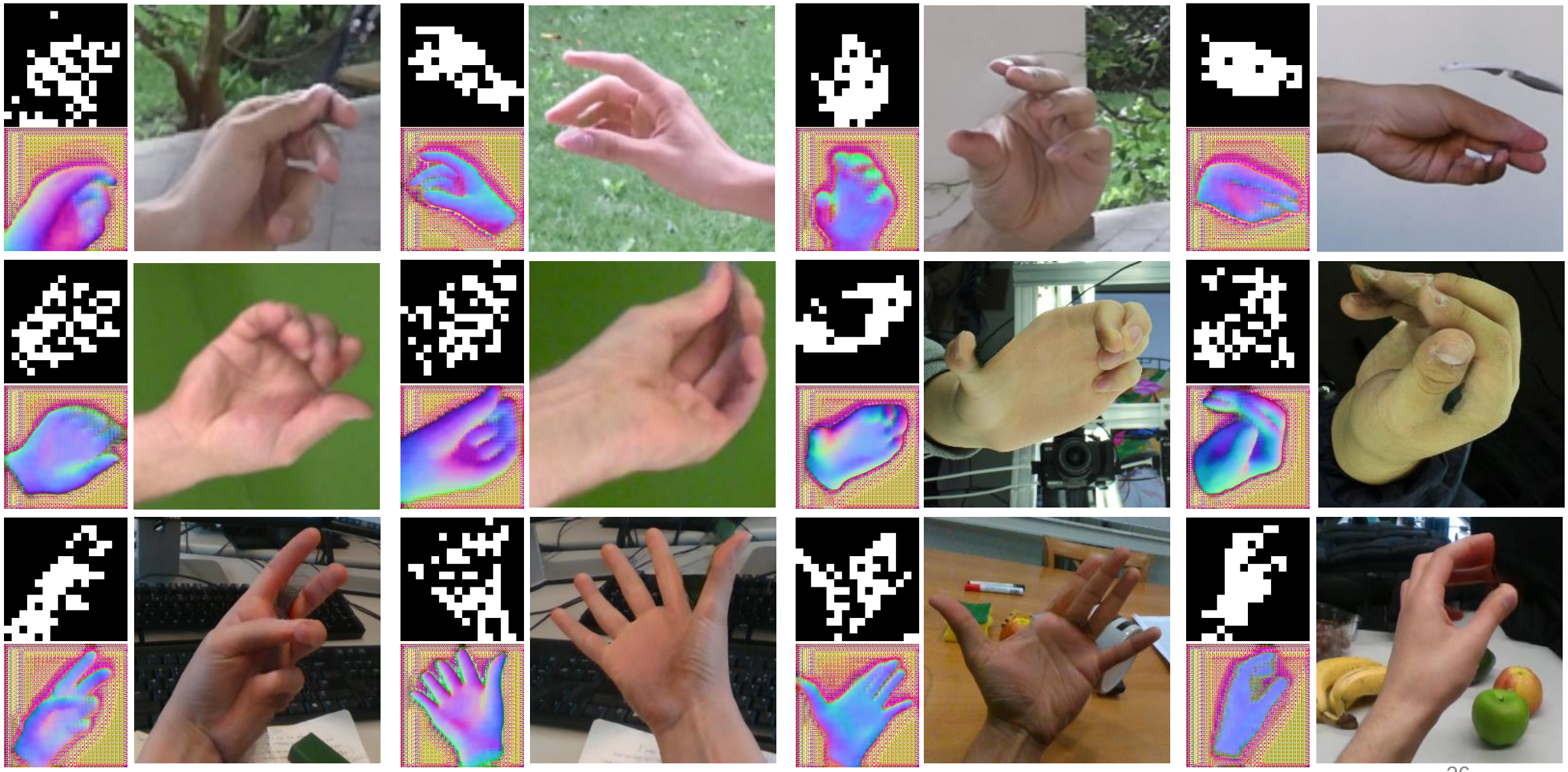






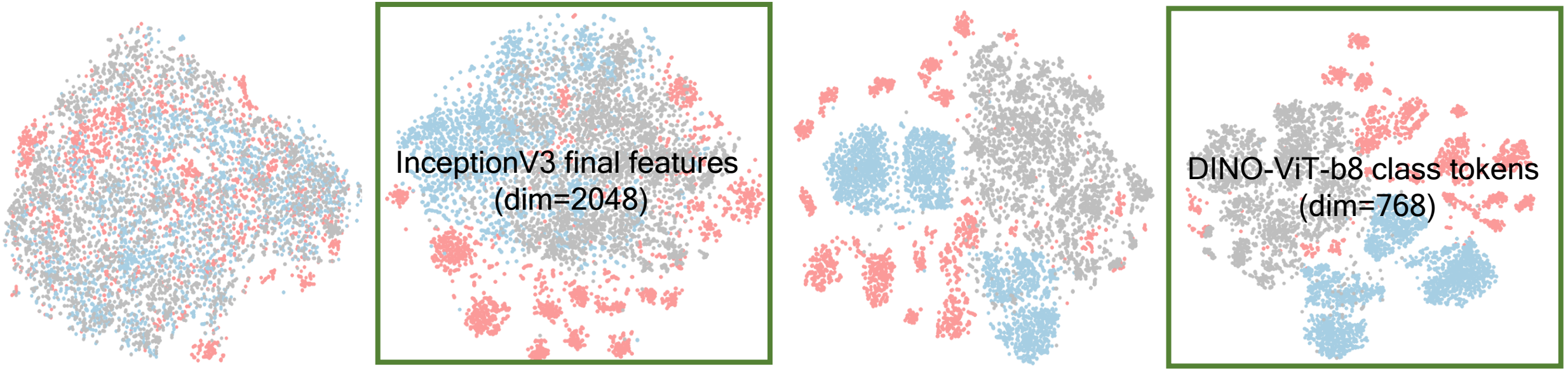
Qualitative Results on Translation $\mathbf{A_2 \rightarrow B}$



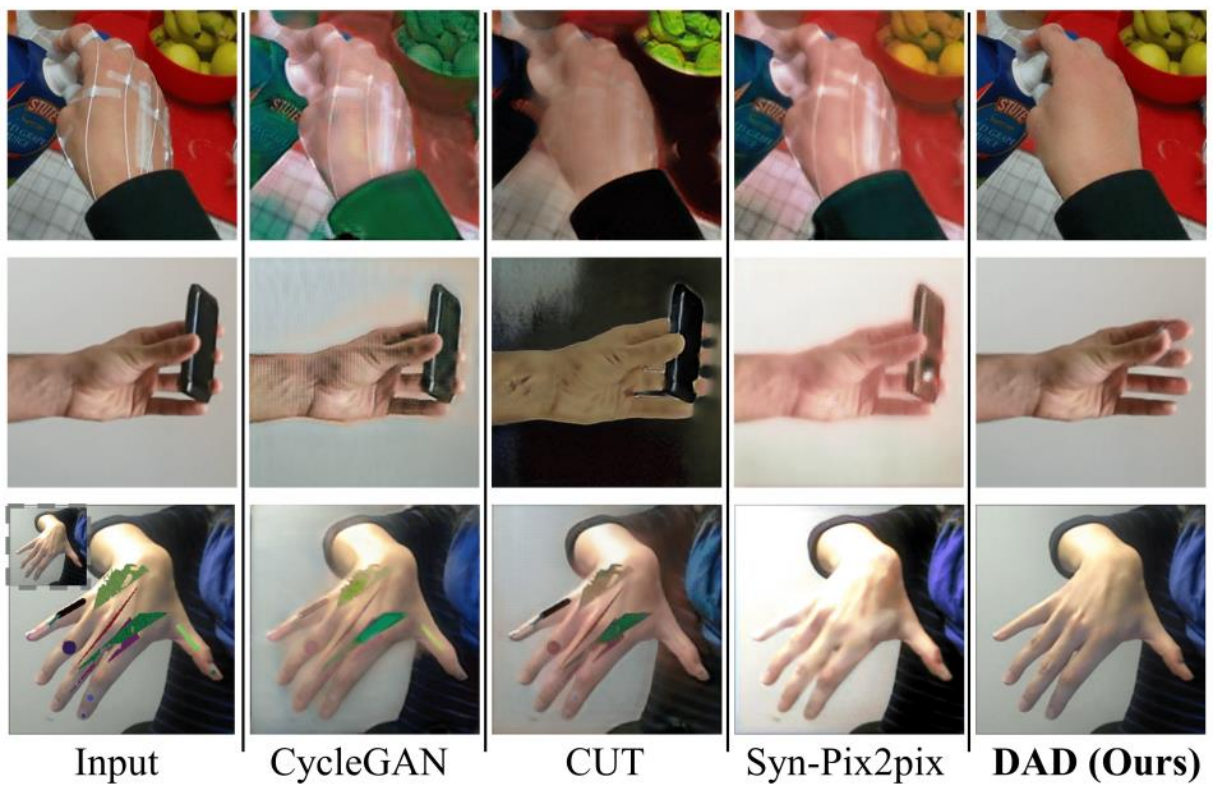
Quantitative Comparisons

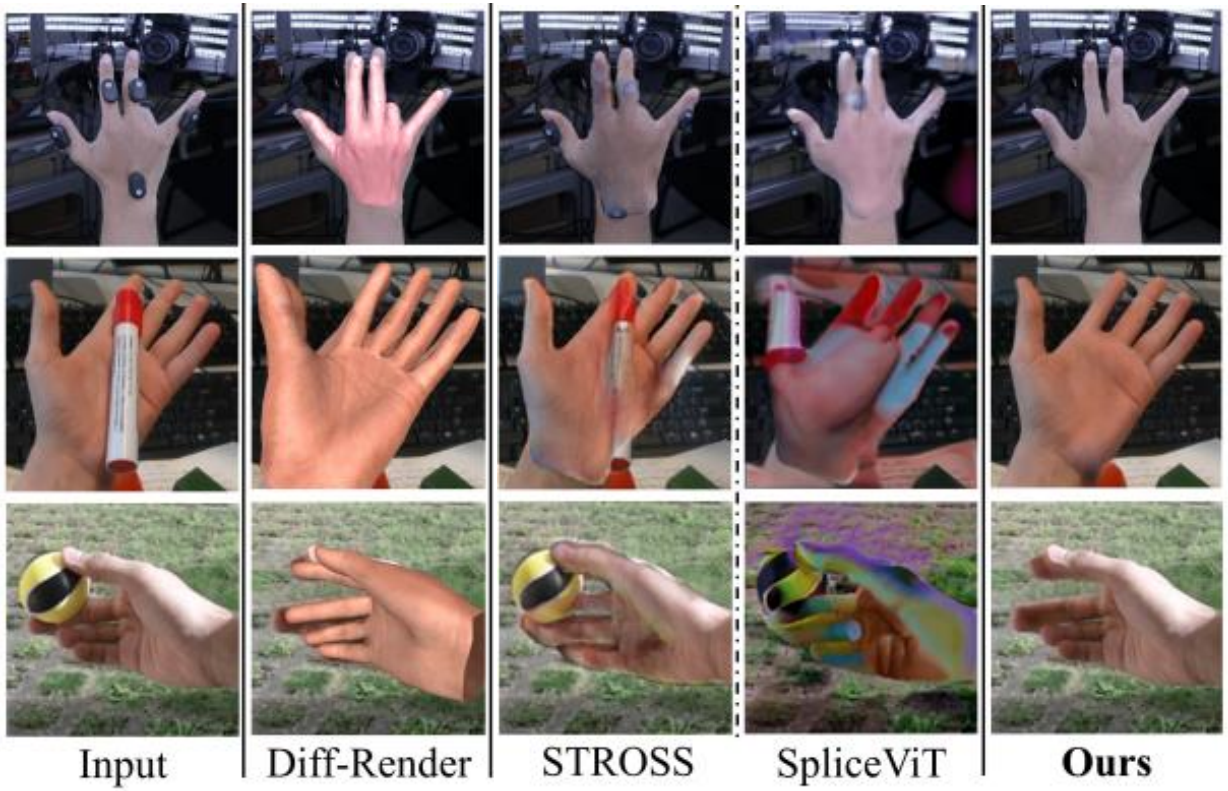
Tasks	$A_1 \rightarrow B$				$A_2 \rightarrow B$			
Metrics	$FID_i \downarrow$	$KID_i(*100) \downarrow$	$FID_v \downarrow$	$KID_v \downarrow$	$FID_i \downarrow$	$KID_i(*100) \downarrow$	$FID_v \downarrow$	$KID_v \downarrow$
CycleGAN [91]	76.39	4.46 ± 0.176	1266.17	32.02 ± 0.994	65.12	4.44 ± 0.196	1021.14	27.13 ± 0.948
GANerated [52]	76.97	4.72 ± 0.153	1220.53	31.72 ± 0.907	68.50	5.05 ± 0.199	985.82	26.75 ± 0.987
H-GAN [56]	93.02	6.53 ± 0.209	1488.32	37.85 ± 1.061	62.94	4.12 ± 0.197	876.94	24.61 ± 0.816
UAG [5]	87.80	5.90 ± 0.177	1375.55	35.67 ± 0.949	70.98	5.35 ± 0.210	1069.45	28.31 ± 0.926
CUT [57]	78.02	5.54 ± 0.192	1230.67	33.34 ± 1.015	58.88	3.73 ± 0.160	749.22	20.02 ± 0.826
Ours	**60.37**	$\mathbf{3.45 \pm 0.236}$	**994.67**	$\mathbf{28.67 \pm 0.916}$	**41.53**	$\mathbf{3.37 \pm 0.154}$	**673.43**	$\mathbf{15.72 \pm 1.209}$

data in $\mathbf{A_1}$ data in $\mathbf{A_2}$ data in $\mathbf{B}$



Qualitative Comparisons





Conclusion

- This work pioneers a semi-supervised image-to-image translation to recover the hand appearance that was originally degraded during the marker-based MoCap process.
- The prior-based sketcher can robustly disentangle the bare structure maps from degraded hand images.
- The DAD is more efficient than unsupervised schemes, and more generalizable than a supervised scheme trained with synthetic degradation.

Project Page