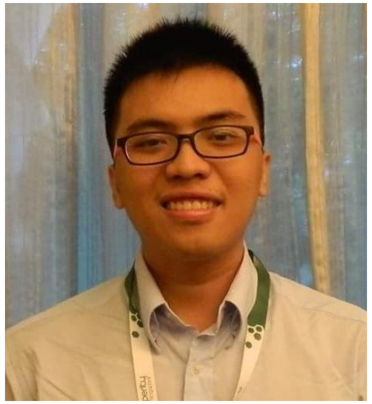
Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition



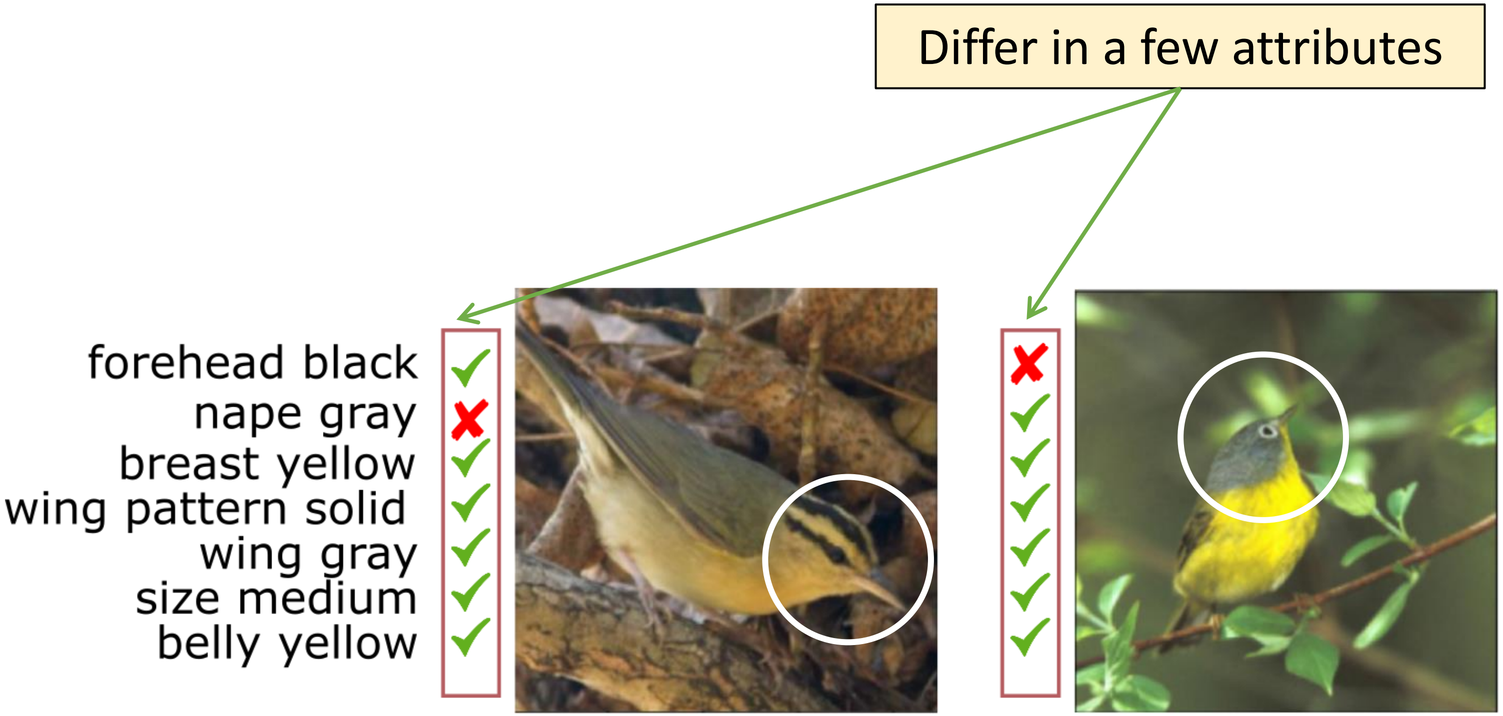
Dat Huynh and **Ehsan Elhamifar**

Khoury College of Computer Sciences

Northeastern University

Motivation

- **Fine-Grained Recognition**: recognize visually **similar classes**
  - Classes **differ in a few attributes**
  - **Costly:** require expert annotator
  - **Cannot handle** unseen classes



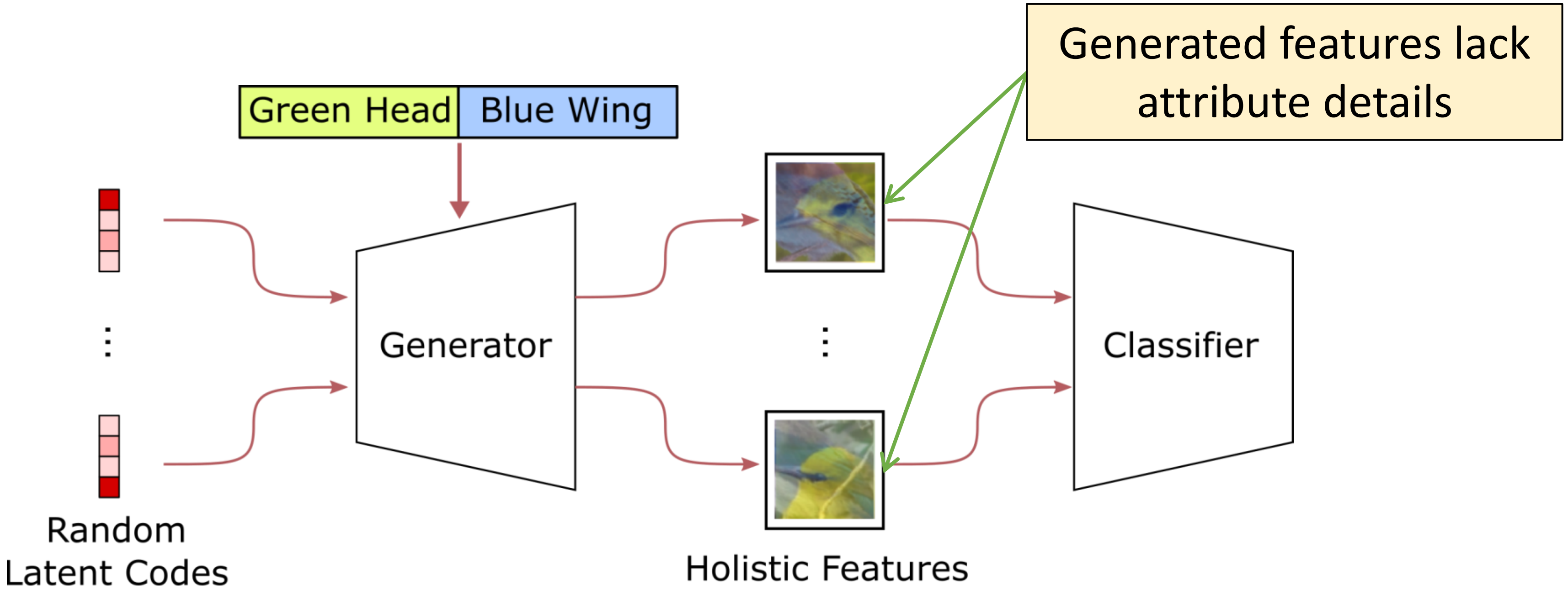
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **Fine-Grained Recognition**: recognize visually **similar classes**
  - Classes **differ in a few attributes**
  - **Costly:** require expert annotator
  - **Cannot handle** unseen classes
- **Zero-Shot Learning**: **recognize unseen classes** without training samples
  - **Reduce** annotation cost

State of the Art

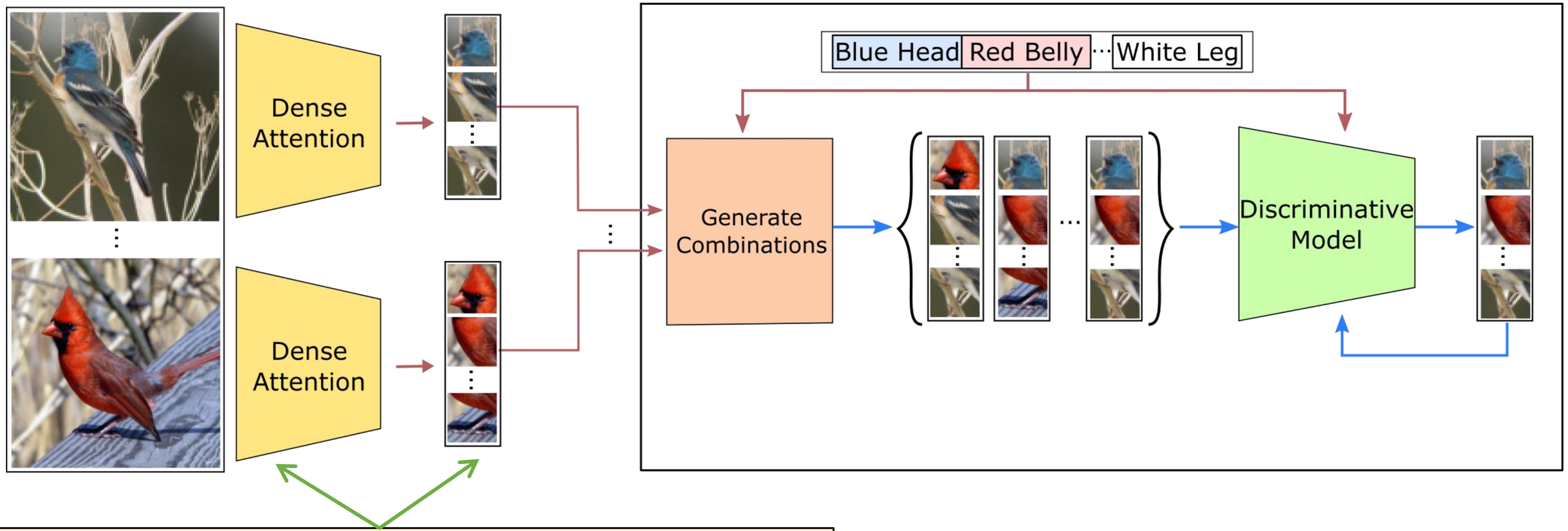
- **Generative Methods:** train a classifier on unseen class features synthesized by a generator
- Challenge:
  - Generate holistic features **lacking attribute details**



[M. Bucher et. al. '17, G. Arora et. al. '18, Y. Xian et. al. '19, E. Schönfeld et. al. ‘19, H. Huang et. al. '19, H. Yu et. al. '19, J. Ni et. al. '19]

Contributions

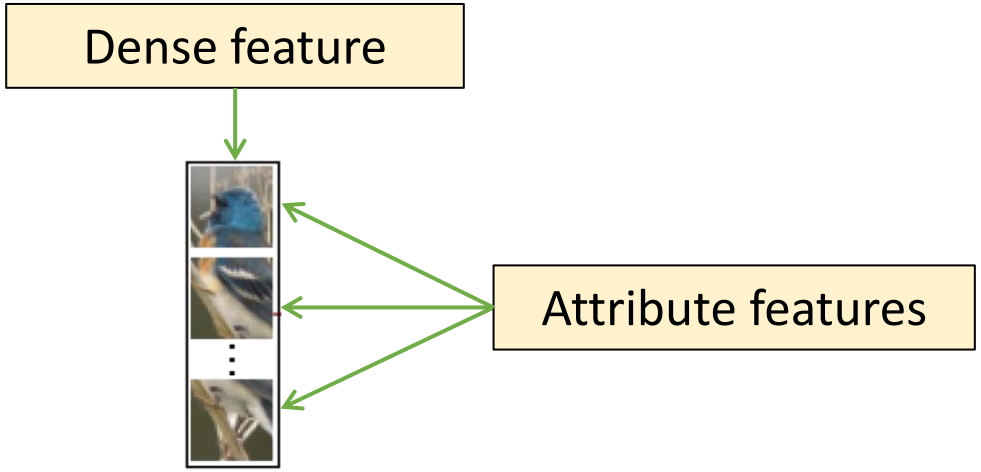
- Address fine-grained ZSL
- Propose a *compositional feature learning* framework:
  - Generate *dense features* **preserving fine-grained details**
  - **Directly** train a classifier **without learning separate generative models**
  - **Outperform SOTA** on DeepFashion, AWA2 and CUB



Extract one feature per attribute
[D. Huynh and E. Elhamifar, CVPR20]

Challenges

- Generate **dense feature** $\boldsymbol{H}$: collection of **attribute features**



- Challenge: **Difficult** to learn generative models
  - **High dimension**
  - **No sample** for unseen classes

Feature Composition

- **Transform** a discriminative model $p(y|\boldsymbol{H}, \boldsymbol{z})$ into a generative model

Bayes Rule

$$\operatorname{argmax}_{\boldsymbol{H}} p(\boldsymbol{H}|y, \boldsymbol{z}) = \operatorname{argmax}_{\boldsymbol{H}} p(y|\boldsymbol{H}, \boldsymbol{z}) p(\boldsymbol{H}|\boldsymbol{z})$$

Find most probable feature for unseen classes

Generative model

Discriminative model

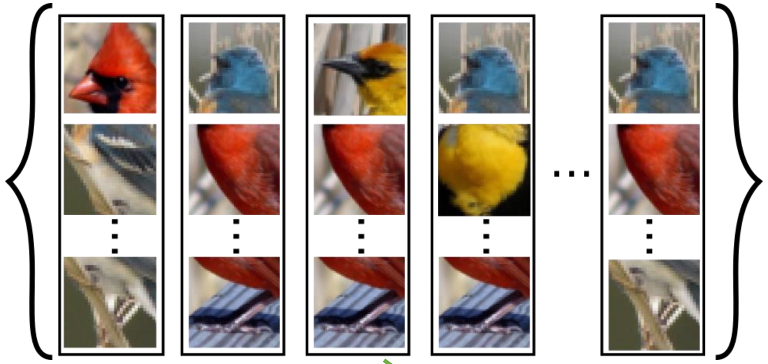
Prior

- Challenge:
  - Search for most probable feature in high dimension space of $\boldsymbol{H}$ is **intractable**

Feature Composition

- Limit our search in combinations of **attribute features across samples**
  - Simulate **unseen attribute combinations**
  - **Tractable** due to finite number of combinations

$\mathcal{U} \triangleq$



Limit our search

$$\operatorname{argmax}_{\boldsymbol{H}}\, p(\boldsymbol{H}|y,\boldsymbol{z}) \approx \operatorname{argmax}_{\boldsymbol{H}\in\mathcal{U}}\, p(y|\boldsymbol{H},\boldsymbol{z})p(\boldsymbol{H}|\boldsymbol{z})$$

Search in all combinations is **expensive**

How to construct **feature prior**?

Feature Prior

- **Restrict the search** to combinations from related sample sets Q_u
  - **Set of samples best reconstruct** the unseen class attributes

Attributes reconstruction loss

$$\mathrm{Q}_u \triangleq \operatorname{argmin}_S \left(\min_{\gamma} \|\boldsymbol{z} - \sum_{i \in S} \boldsymbol{z}^i \gamma_i\|_2^2 \right)$$

- **Construct feature prior**
  - **Independence** among attribute features
  - The **more related** a sample, the **more probable** its feature will be used

Measure sample relatedness

$$p(\boldsymbol{H}|\boldsymbol{z}^u) \triangleq \prod_{a=1}^{A} p(\boldsymbol{h}^a_{i_a}|\boldsymbol{z}^u), \quad p(\boldsymbol{h}^a_{i_a}|\boldsymbol{z}^u) \triangleq \begin{cases} \dfrac{\exp\left(\langle \boldsymbol{z}^{i_a}, \boldsymbol{z}^u \rangle\right)}{\sum_{i \in \mathrm{Q}_u(S)} \exp\left(\langle \boldsymbol{z}^i, \boldsymbol{z}^u \rangle\right)}, & \text{if } i_a \in \mathrm{Q}_u, \\ 0, & \text{otherwise}, \end{cases}$$

Attribute feature of sample i_a

Proposed Method

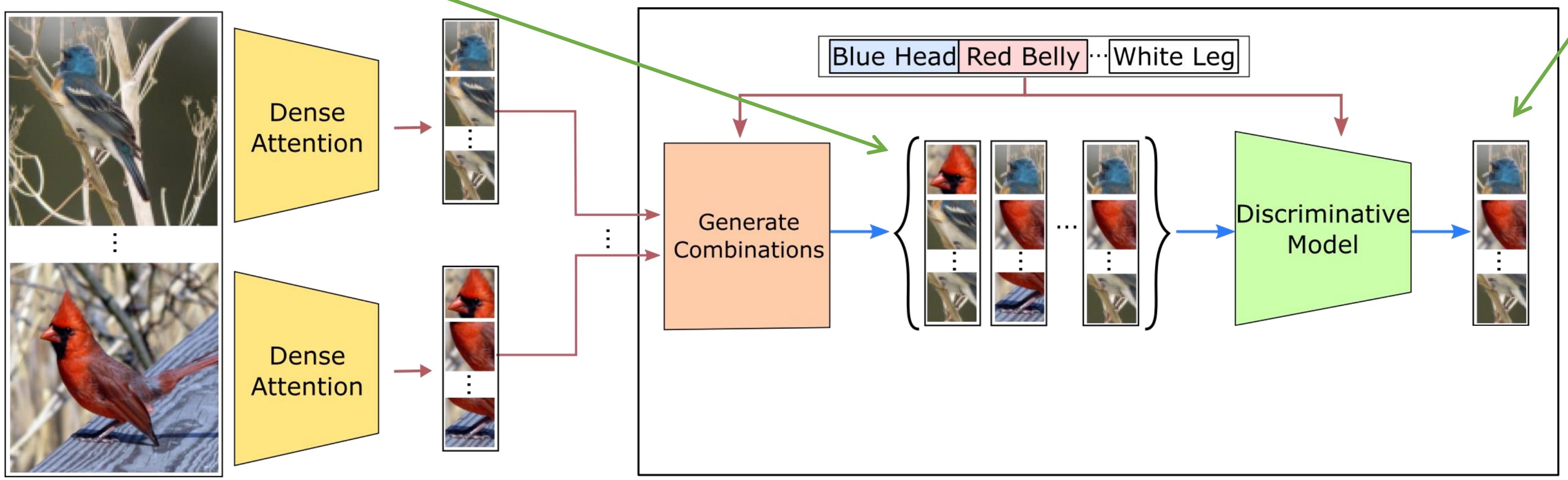
1. **Generate candidate combinations** from feature prior
2. **Pick the most probable combination** among candidates

$$M_u = \{\boldsymbol{H} | \boldsymbol{H} \sim p(\boldsymbol{H}|\boldsymbol{z}^u)\}$$

$$\boldsymbol{H}_u \triangleq \mathrm{argmax}_{\boldsymbol{H} \in M_u} \, p(u|\boldsymbol{H}, \boldsymbol{z}^u) p(\boldsymbol{H}|\boldsymbol{z}^u)$$

generate candidate combinations

Select most probable combination



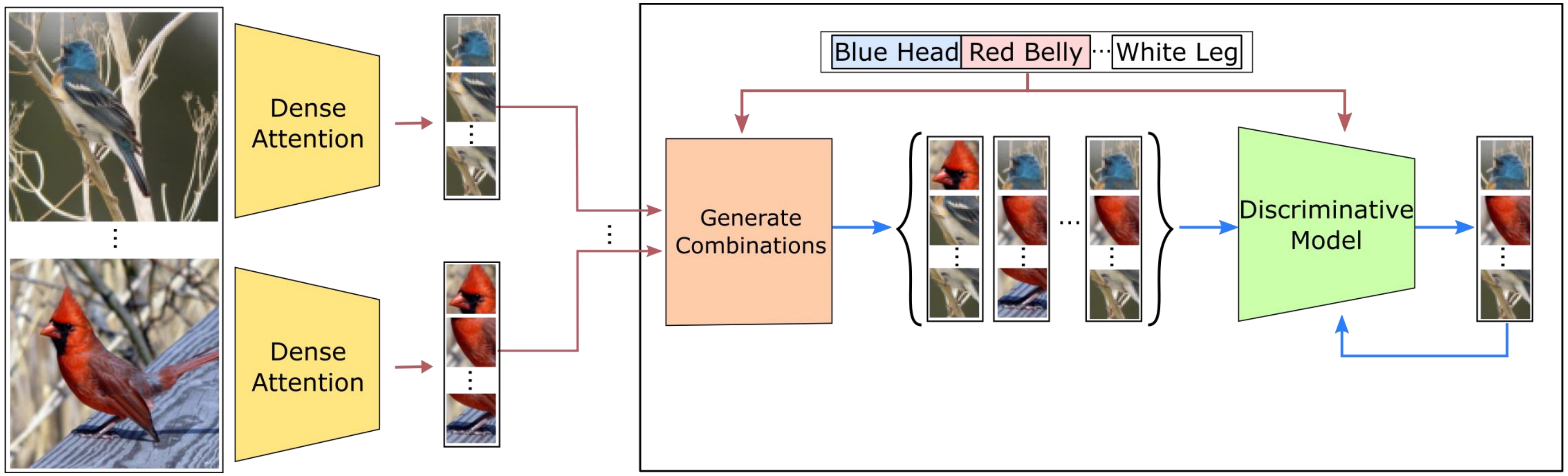
Proposed Method

1. **Generate candidate combinations** from feature prior
2. **Pick the most probable combination** among candidates
3. **Train the discriminative model** on real and composed features

$$\min \mathbb{E}_S \left[-\frac{1}{|S|} \sum_{i \in S} y_i \log p\big(y_i | \boldsymbol{H}_i\, \boldsymbol{z}^{y_i}\big) - \frac{1}{|\mathcal{C}_u|} \sum_{u \in \mathcal{C}_u} u \log p\big(u | \boldsymbol{H}_u, \boldsymbol{z}^u\big) \right]$$

Cross-entropy with seen class features

Cross-entropy with composed features



Experiments

- **Outperform SOTA** on DeepFashion, AWA2, CUB

Method	**DFashion** (5691 images/class)				**AWA2** (588 images/class)				**CUB** (47 images/class)			
	$u \to u$	$a \to s$	$a \to u$	H	$u \to u$	$a \to s$	$a \to u$	H	$u \to u$	$a \to s$	$a \to u$	H
CVC	-	-	-	-	71.1	81.4	56.4	66.7	54.4	47.6	47.4	47.5
TripletLoss	-	-	-	-	67.9	**83.2**	48.5	61.3	63.8	52.3	55.8	53.0
f-VAEGAN-d2	-	-	-	-	71.1	70.6	57.6	63.5	61.0*	60.1*	48.4*	53.6*
CADA-VAE	-	-	-	-	-	75.0	55.8	64.0	-	53.5	51.6	52.5
f-Translator	40.7	30.5	23.9	26.8	70.4	72.6	55.3	62.6	58.5	54.8	47.0	50.6
DAZLE	38.4	**38.1**	21.5	27.5	67.9	75.7	60.3	67.1	65.9	**59.6**	56.7	58.1
Composer (Ours)	**43.0**	32.9	**31.2**	**32.0**	**71.5**	77.3	**62.1**	**68.8**	**69.4**	56.4	**63.8**	**59.9**

+4.5%

+1.7%

+1.8%

Qualitative Results

- Generated features are **interpretable**