KOREA UNIVERSITY

LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data

CVPR 2023

Jihye Park*, Sunwoo Kim*, Soohyun Kim*,
Seokju Cho, Jaejun Yoo, Youngjung Uh, Seungryong Kim
Korea University

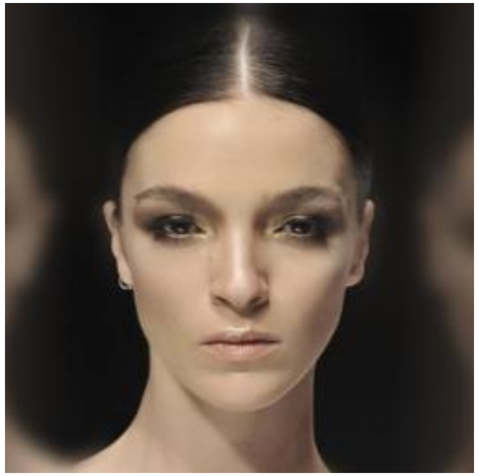
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

What is Image-to-Image Translation

Content Image

Reference Image/Text

Synthesized Image



“ocean”
“sunset”
“cloudy”

Motivation and Problem Formulation

Per-sample-level

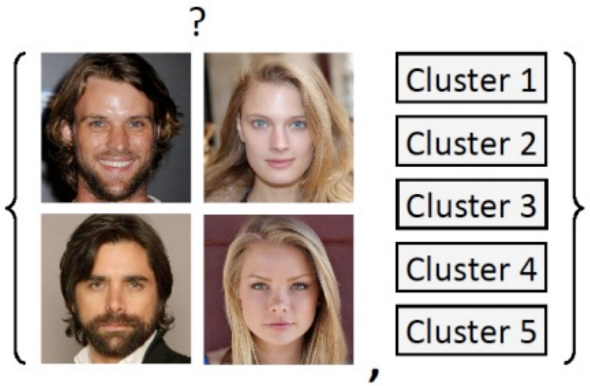
- <u>Each image is annotated corresponding classes through **manual labeling**</u>

✗ *Intensive labeling resource*

StarGANv2 (CVPR'20)

Unsupervised

- <u>Pseudo labels from **clustering or SSL of unsupervised manner**</u>



✓ *No ground-truth labels*

TUNIT (ICCV'21)
Styla-aware Discriminator. (CVPR'22)

Motivation and Problem Formulation

Clusters do not have semantic meaning

Designated candidate domains have semantic meanings

Motivation and Problem Formulation

Per-sample-level

- <u>Each image is annotated corresponding classes through **manual labeling**</u>

Face

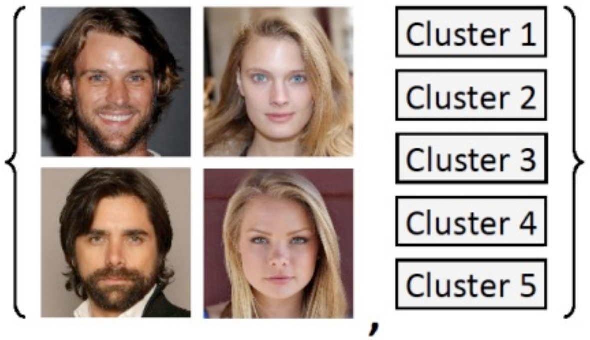
✗ *Intensive labeling resource*

✓ *Applicable: we can accurately target the attributes to synthesize*

Unsupervised

- <u>Pseudo labels from **clustering or SSL of unsupervised manner**</u>

?



✓ *No ground-truth labels*

✗ *Not applicable: we can't accurately target the attributes to synthesize*

Dataset-level (Ours)

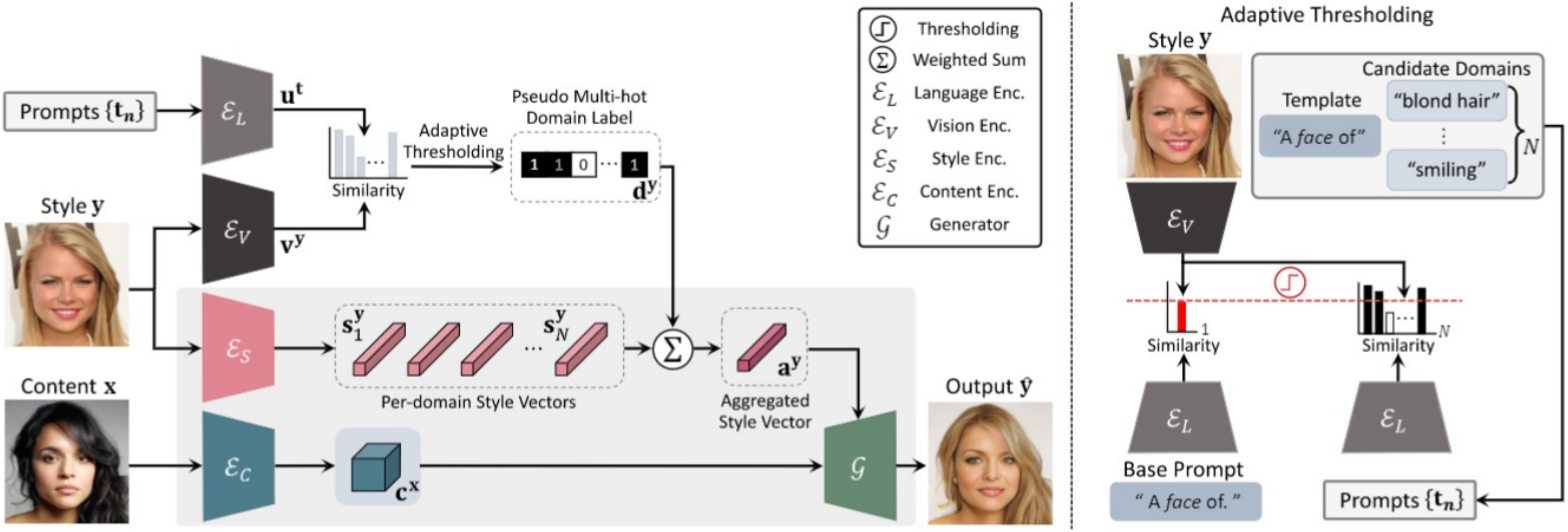
- <u>Pseudo labels from **Weak human's supervision** and **large VL model**</u>

Face

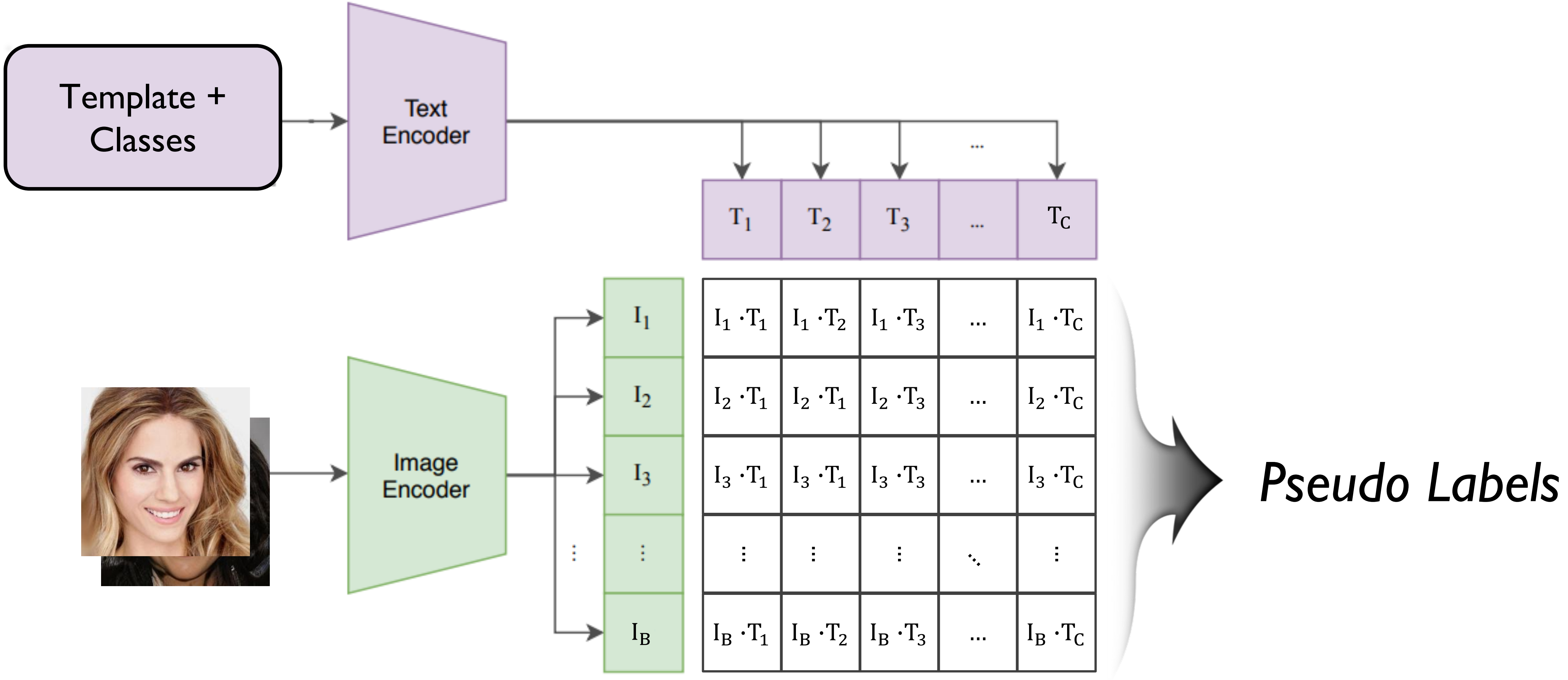
✓ *No ground-truth labels*

✓ *Applicable: we can accurately target the attributes to synthesize*

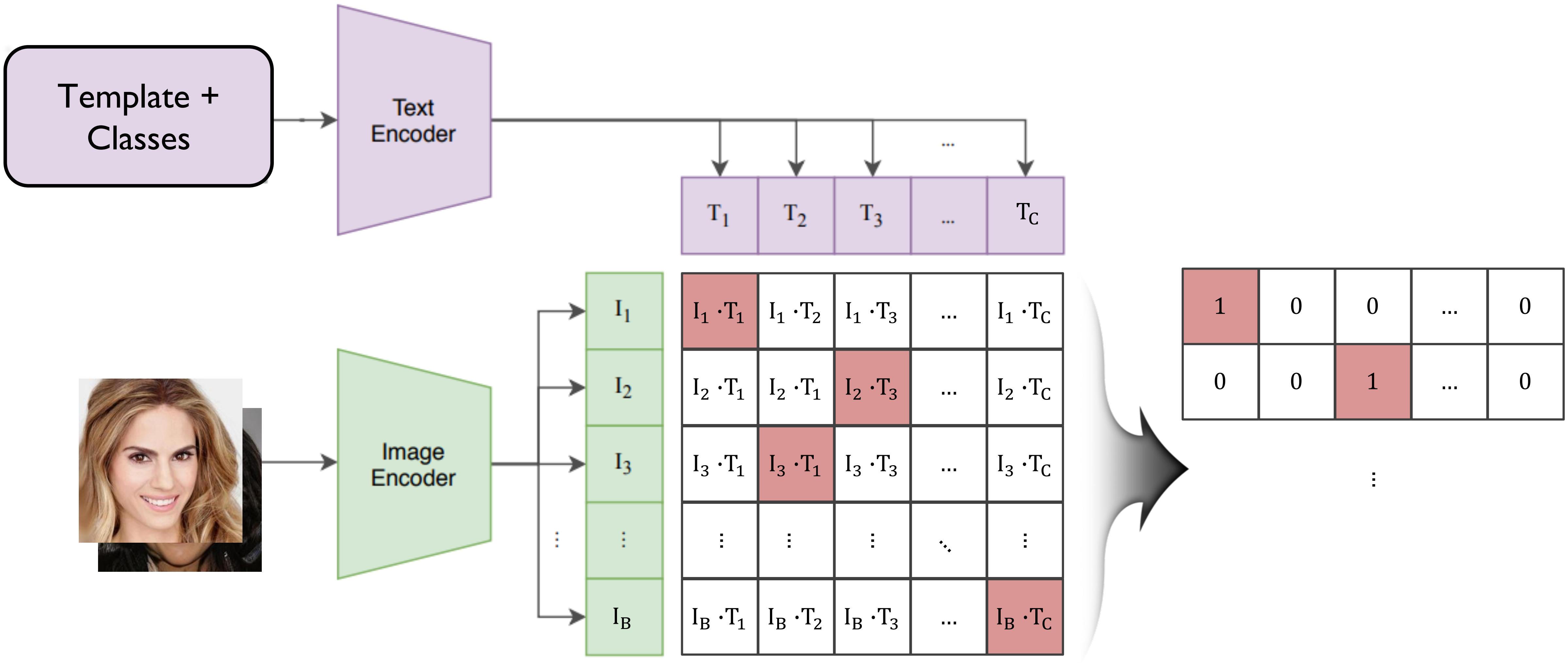
Proposed Network



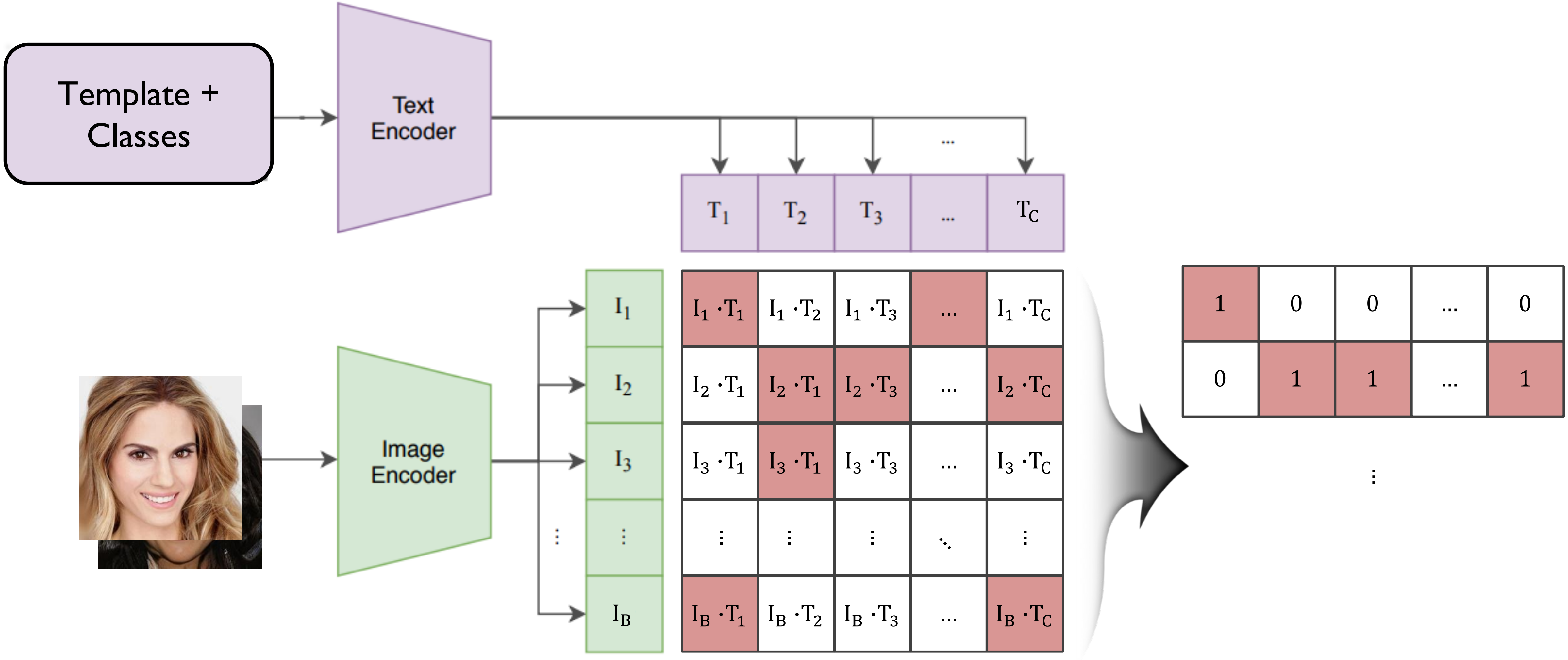
How to make pseudo labels?



How to make pseudo labels?: **Topk (k=1)**

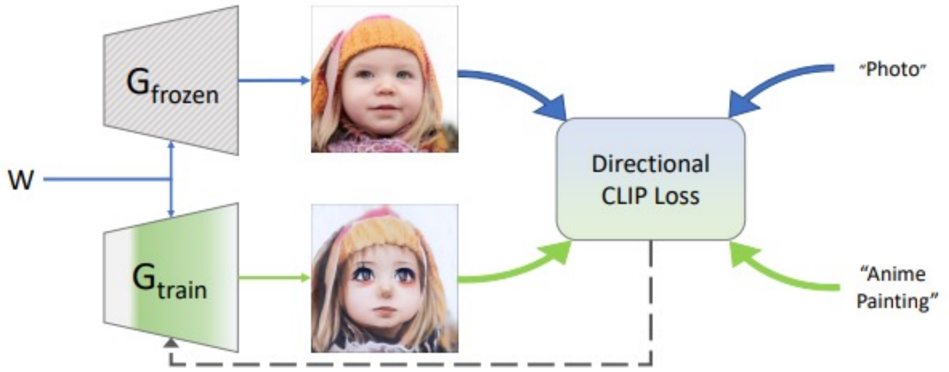


How to make pseudo labels?: **Thres (τ=0.2)**



How to make pseudo labels?: **Adaptive Thres**

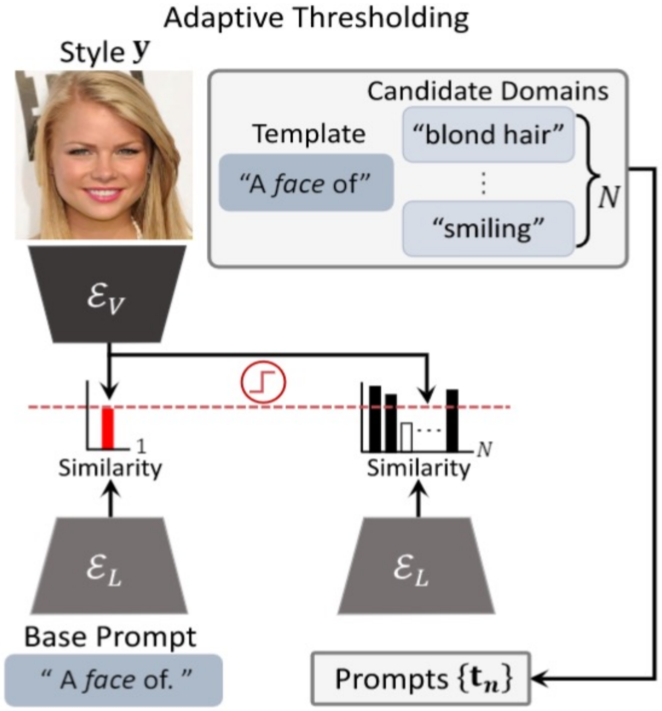
Direction cliploss



$$\Delta T = E_T\left(t_{target}\right) - E_T\left(t_{source}\right) \,,$$
$$\Delta I = E_I\left(G_{train}\left(w\right)\right) - E_I\left(G_{frozen}\left(w\right)\right) \,,$$
$$\mathcal{L}_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I|\,|\Delta T|} \,.$$

StyleGAN-NADA (SIGGRAPH'22)

Adaptive Thresholding

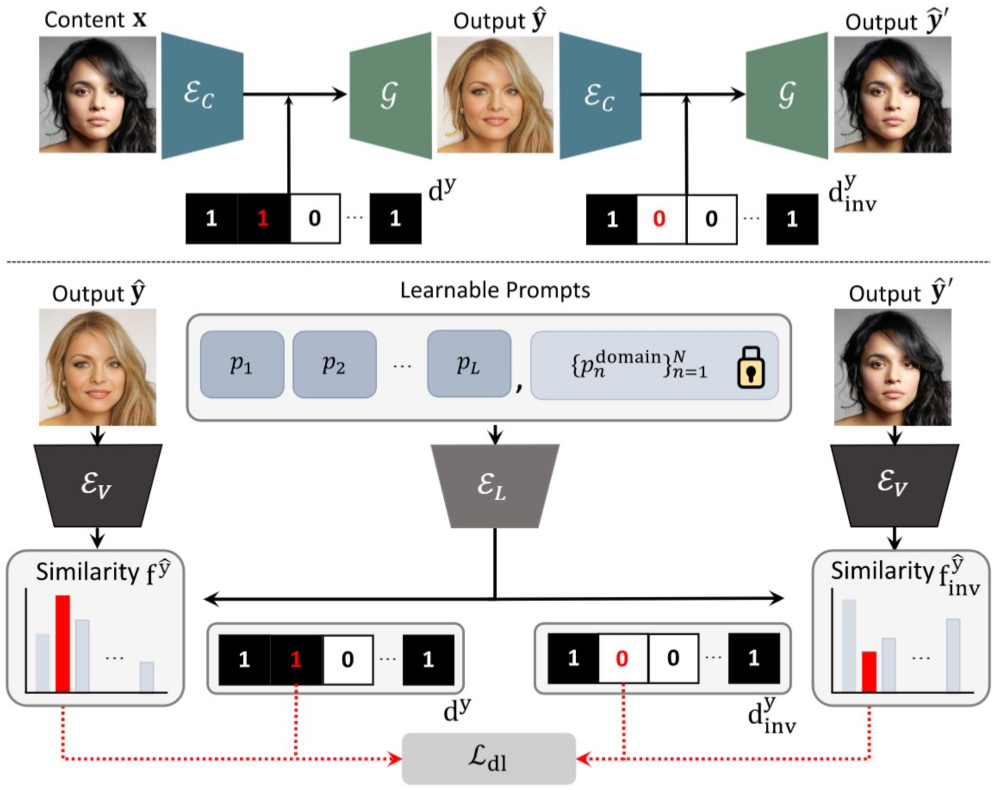


How to improve performance of CLIP?

Template Augmentation

```
[
    ['a face photo with.'],
    ['a face photo of the.'],
    ['the face photo of the.'],
    ['a good face photo of the.'],
    ["high quality face photo of."],
    ["a face image of."],
    ["the face image of."],
    ["high quality face image of."],
    ["a high quality face image of."],
]
```

Prompt leanring



Loss Functions

Adversarial Loss($\mathcal{L}_{adv}$)

$$= \mathbb{E}_{\mathbf{x},\mathbf{y}} \sum_{n=1}^{N} [\log \mathcal{D}_n(\mathbf{y}) d_n^{\mathbf{y}} + \log(1 - \mathcal{D}_n(\mathcal{G}(\mathbf{x}, \mathbf{a}^{\mathbf{y}})) d_n^{\mathbf{y}})]$$

Domain Regularization Loss($\mathcal{L}_{\mathrm{dl}}$)

$$\mathcal{L}_{\mathrm{dl}} = \mathcal{H}(d_n^{\mathbf{y}}, f_n^{\hat{\mathbf{y}}}) + \mathcal{H}(d_{\mathrm{inv},n}^{\mathbf{y}}(n), f_{\mathrm{inv},n}^{\hat{\mathbf{y}}})$$

Cycle-Consistency Loss($\mathcal{L}_{\mathrm{cyc}}$)

$$\mathcal{L}_{\mathrm{cyc}} = \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\|\mathbf{x} - \mathcal{G}(\mathbf{c}^{\hat{\mathbf{y}}}, \mathbf{a}^{\mathbf{x}})\|_1 \right]$$

Style Reconstruction Loss($\mathcal{L}_{\mathrm{sty}}$)

$$\mathcal{L}_{\mathrm{sty}} = \mathbb{E}_{\mathbf{x},\mathbf{y}} [\|\mathbf{s}^{\mathbf{y}} - \mathcal{E}_S(\hat{\mathbf{y}})\|_1]$$

Overall Objective($\mathcal{L}_{\mathrm{total}}$)

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{dl}} \mathcal{L}_{\mathrm{dl}} + \lambda_{\mathrm{cyc}} \mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{sty}} \mathcal{L}_{\mathrm{sty}}$$

n'th Discriminator: $\mathcal{D}_n(\cdot)$

Binary Cross Entropy: $\mathcal{H}(\cdot, \cdot)$

Quantitative Results

Our proposed techniques improve the performance of CLIP in our framework !

	AnimalFaces-10 [6]					CelebA-HQ [7]				
N	Top-1	Top-3	Baseline	TextAug	Prompt learning	Top-1	Top-3	Baseline	TextAug	Prompt learning
4	0.762	0.672	0.678	0.796	**0.832**	0.372	0.421	0.423	0.435	**0.481**
7	0.903	0.701	0.688	0.862	**0.893**	0.423	0.613	0.610	0.631	**0.652**
10	0.956	0.723	0.693	0.835	**0.880**	0.355	0.562	0.610	0.638	**0.670**
13	0.826	0.654	0.606	0.785	**0.801**	0.293	0.533	0.591	0.612	**0.639**
16	0.753	0.630	0.601	0.753	**0.783**	0.300	0.522	0.562	0.613	**0.641**

Table 4. **F1 score on the varying the number of domains.**

Quantitative Results

Comparison to other works

Method	CelebA-HQ [41] mFID	CelebA-HQ [41] D&C	AnimalFaces-10 [35] mFID	AnimalFaces-10 [35] D&C	Food-10 [7] mFID	Food-10 [7] D&C
StarGAN2 [10] (sup.)	32.16	1.22 / **0.446**	**33.67**	**1.54** / **0.91**	65.03	1.09 / 0.76
Smoothing [40] (sup.)	35.93	**1.25** / 0.431	38.93	0.97 / 0.75	61.13	0.96 / 0.68
TUNIT [3] (unsup.)	61.29	0.24 / 0.13	47.70	1.04 / 0.81	52.2	1.08 / **0.88**
Kim *et al.* [29] (unsup.)	41.33	0.60 / 0.241	36.83	1.06 / 0.82	**49.34**	1.06 / 0.80
LANIT	**27.96**	0.91/ 0.34	34.11	1.46 / 0.89	49.50	**1.24** / 0.86

Ablation study
(get candidate dominas form dictionary)

Datasets	Default mFID	Default D&C	Dictionary mFID	Dictionary D&C
CelebA-HQ [15]	27.96	0.91/0.34	28.34	0.77/0.23
AnimalFaces-10 [35]	34.11	1.46/0.89	40.48	1.01/0.78
Food-10 [7]	48.08	1.24/0.86	49.50	1.17/0.81

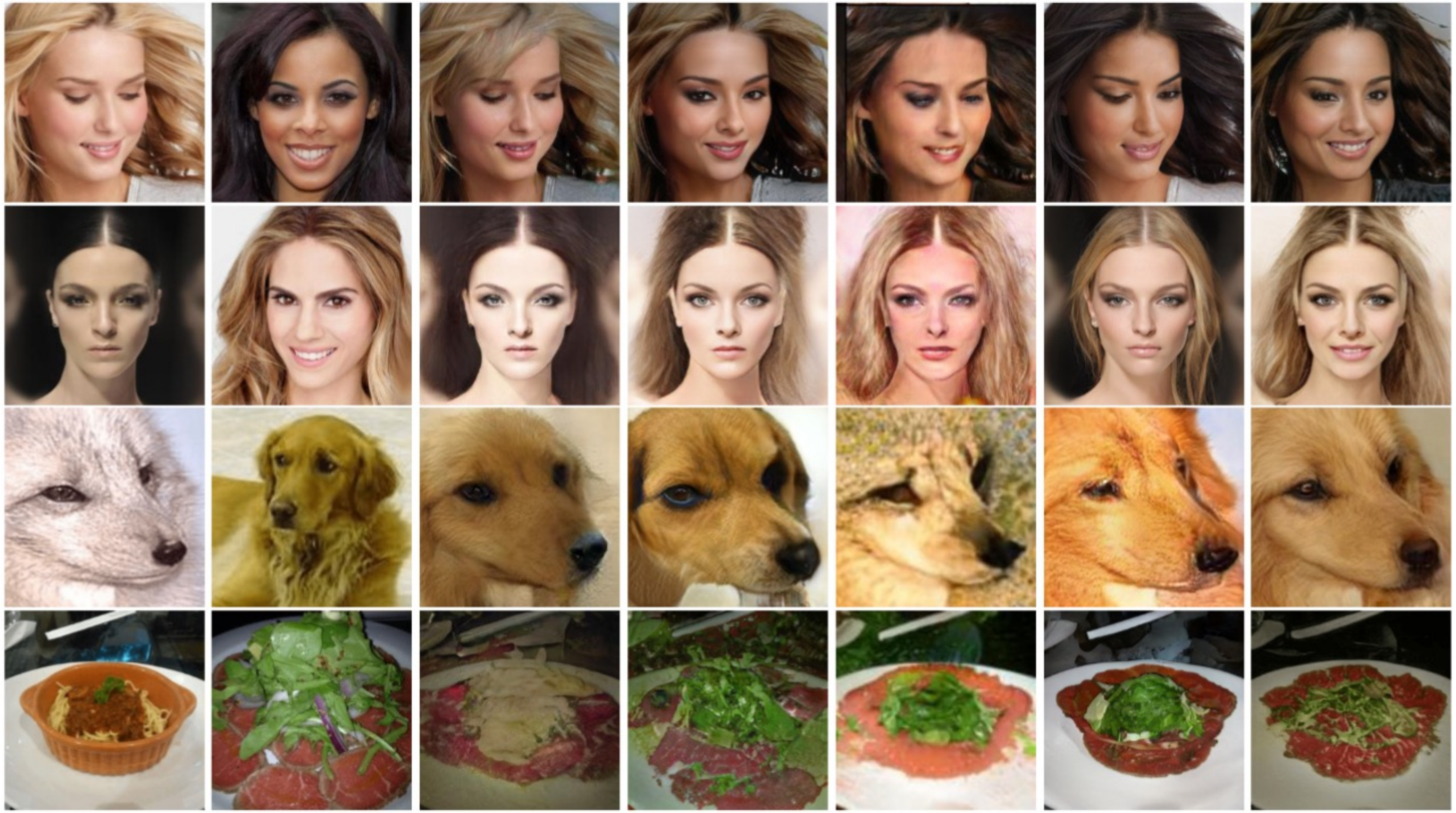
Ablation study(#of domain to train)

N	Method	AnimalFaces-10 [35] mFID	AnimalFaces-10 [35] D&C	CelebA-HQ [41] mFID	CelebA-HQ [41] D&C
4	TUNIT	77.7	0.88 / 0.74	61.5	0.24 / 0.12
	LANIT	71.6	1.35 / 0.46	49.3	0.33 / 0.14
7	TUNIT	62.7	1.02 / 0.73	54.7	0.33 / 0.16
	LANIT	49.9	1.47 / 0.66	43.2	0.44 / 0.19
10	TUNIT	47.7	1.04 / 0.81	61.3	0.24 / 0.13
	LANIT	**34.1**	1.46 / **0.89**	27.96	**0.91** / **0.34**
13	TUNIT	56.8	0.99 /0.72	98.9	0.08 / 0.03
	LANIT	30.13	1.43 / 0.85	34.8	0.58 / 0.21
16	TUNIT	54.1	1.09 / 0.78	127.7	0.04 / 0.02
	LANIT	35.8	**1.49** / 0.82	**27.92**	0.76 / 0.23

Ablation study(components)

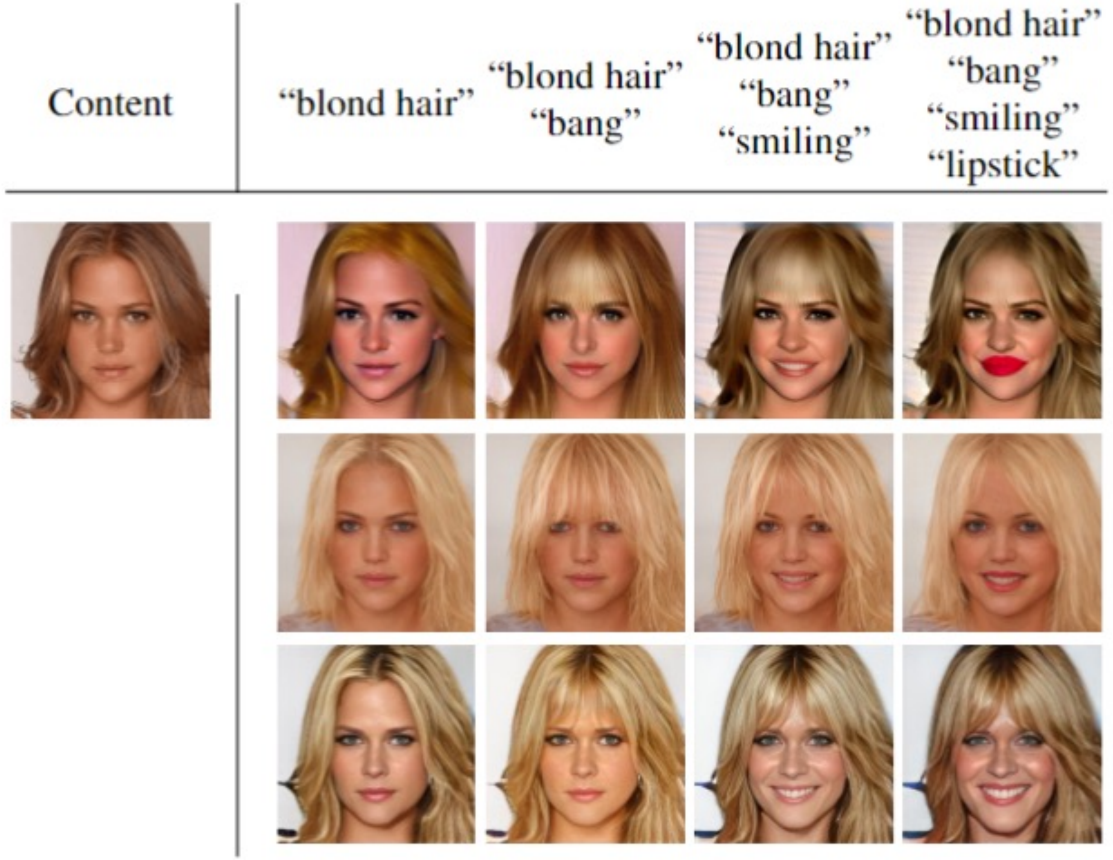
	Method	CelebA-HQ [41] mFID	CelebA-HQ [41] D&C	CelebA-HQ [41] F1.
A	LANIT (Top-1)	49.65	0.56 / 0.32	0.347
B	LANIT (Top-3)	41.68	0.68 / 0.30	0.564
C	Baseline	33.79	0.70 / 0.26	0.613
D	+ DL Loss	29.21	0.86 / 0.32	0.632
E	+ Prompt Learning	**27.96**	**0.91** / **0.34**	**0.718**

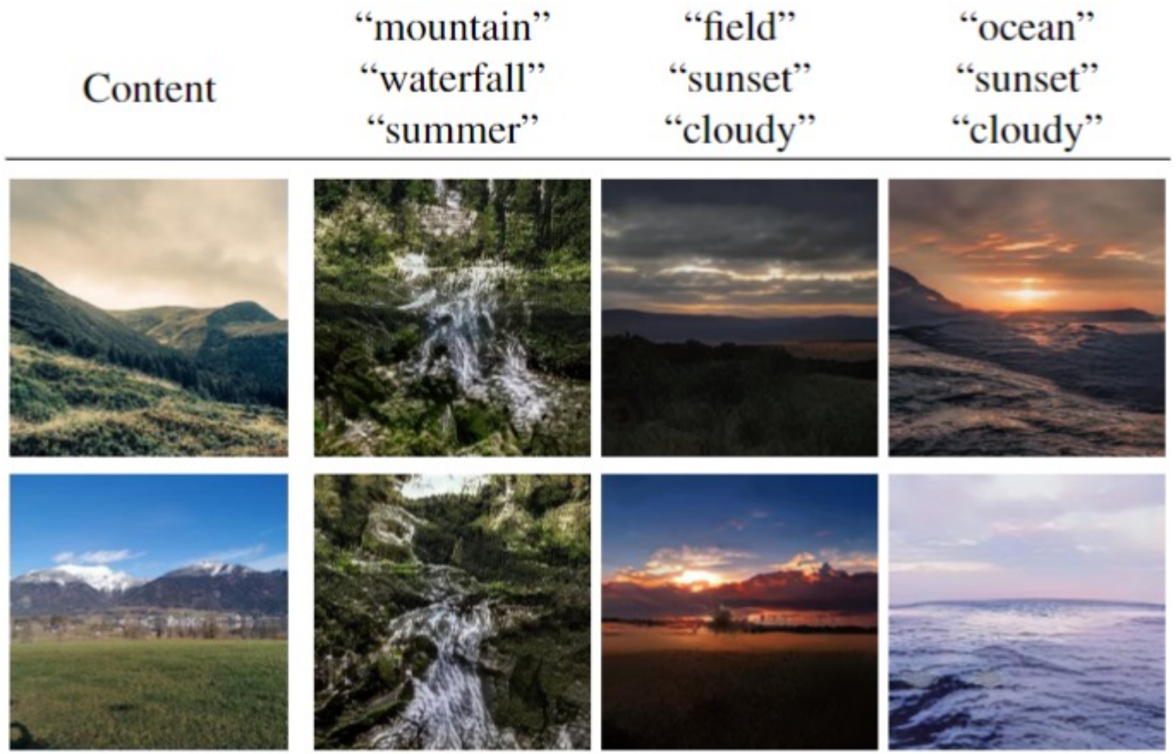
Qualitative Results: Comparison



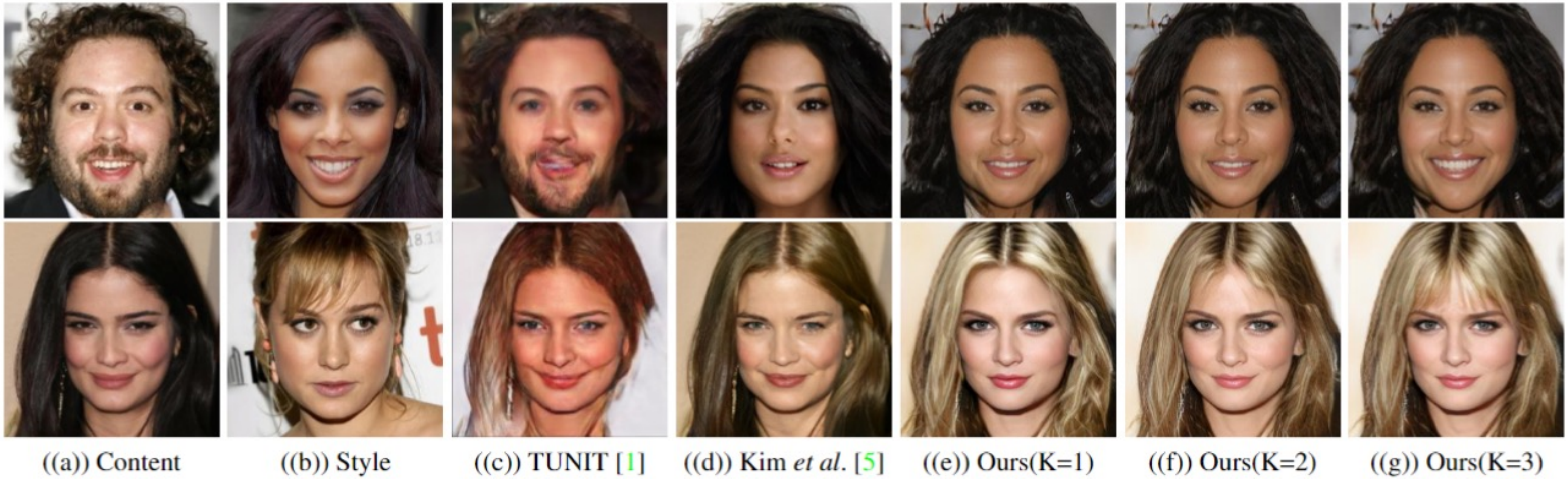
(a) Content (b) Style (c) StarGANv2 [10] (d) Smoothing [40] (e) TUNIT [3] (f) Kim *et al.* [29] (g) Ours

Latent Guided Manipulation





Reference Guided Manipulation



Qualitative Results on Various Datasets

AnimalFaces-10

MetFace

LSUN-Church

Food-10

Anime

LSUN-Car

Thank you!