Image Sentiment Transfer

Tianlang Chen, Wei Xiong, Haitian Zheng, Jiebo Luo

ACM Multimedia 2020 link

20.08.05
발표: 정채연

Contents

1. Motivation
2. Contribution
3. Related Works
4. Methodology
5. Experiments & Results

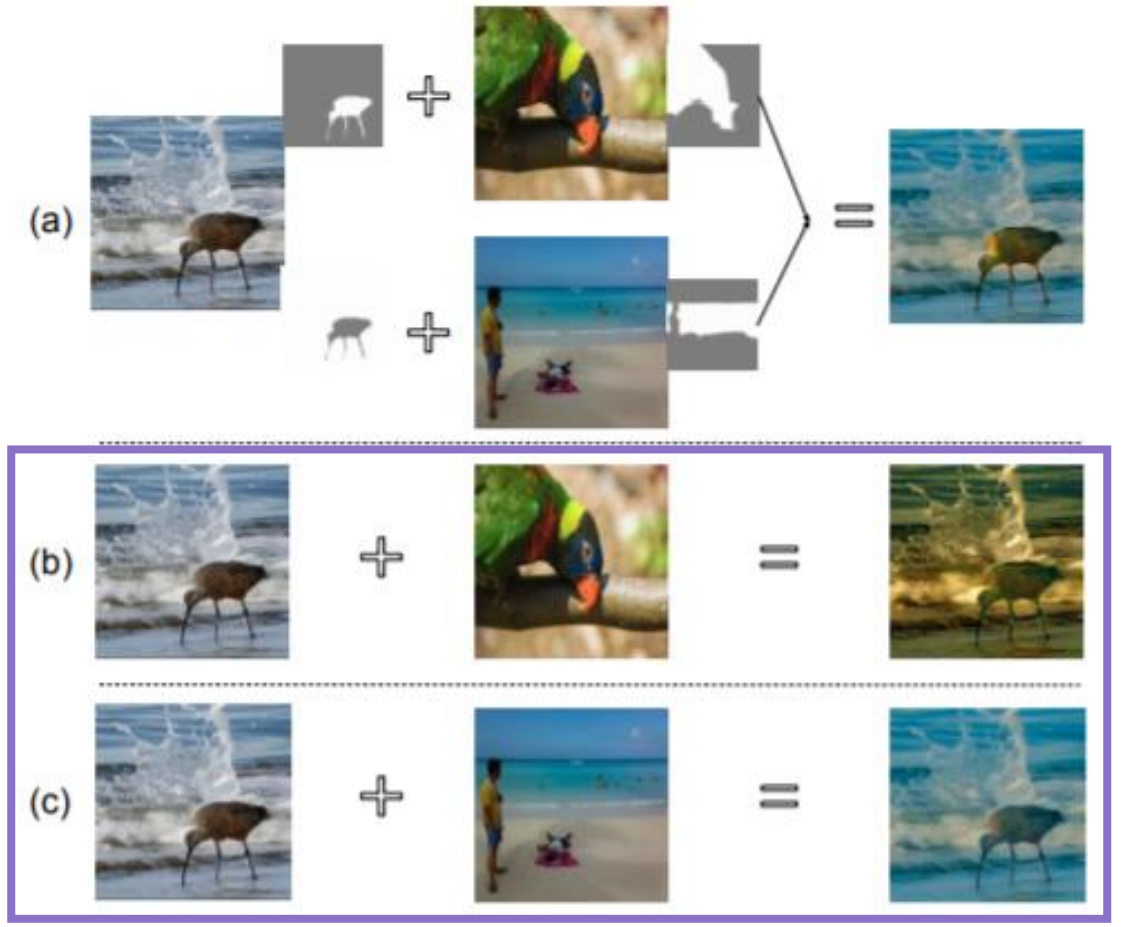
Motivation

Problem

- the rule to transfer the sentiment of each contained object can be completely different
- Image sentiment is more sensitive and related to color-based elements (contrast, saturation, brightness, dominant color)

Solution

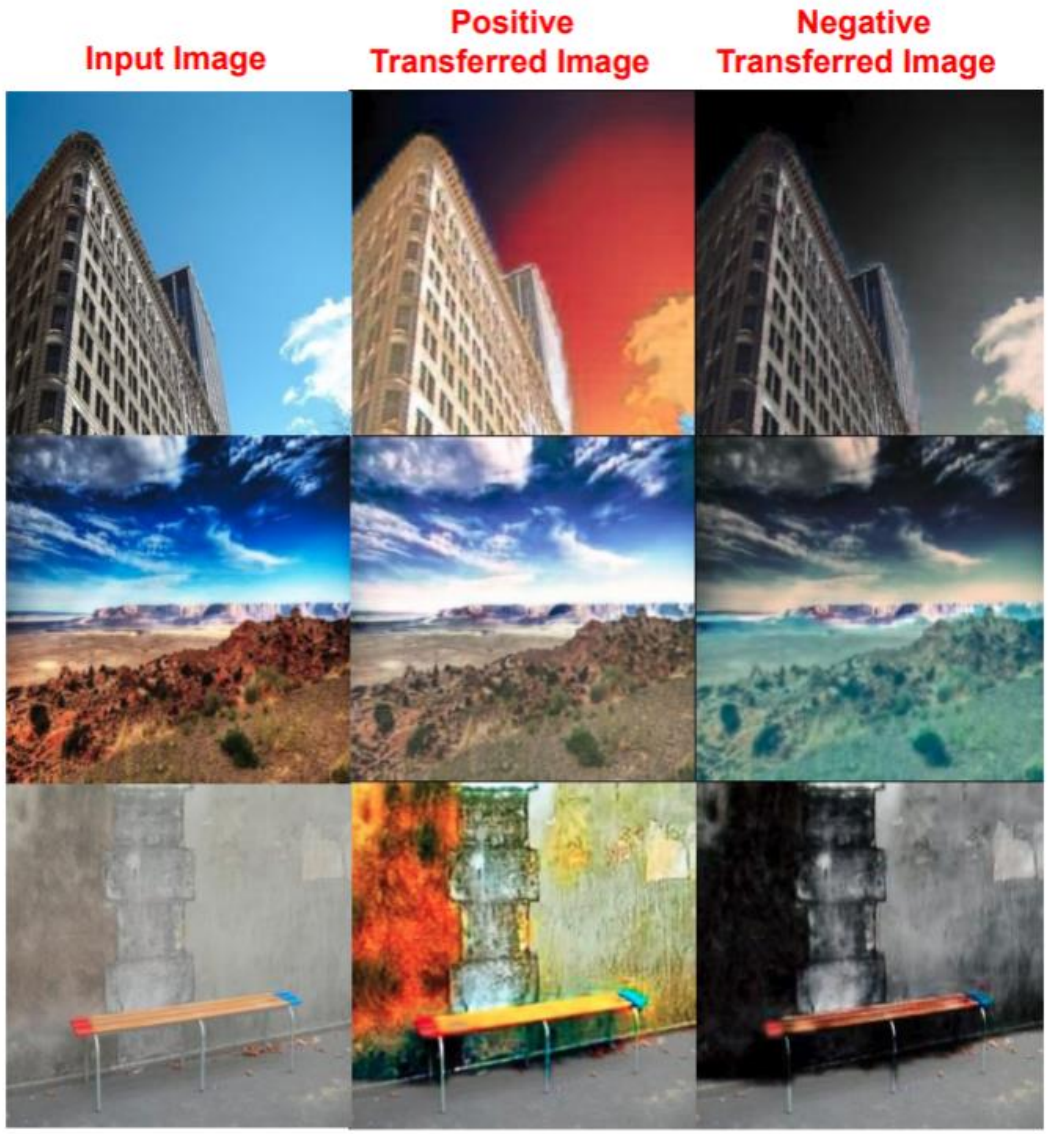
- image sentiment transfer at the object level
- Sentiment-aware GAN (SentiGAN) → transfer object-level, color-based information



global image transfer by a single reference image
→ Inadequate

Contribution

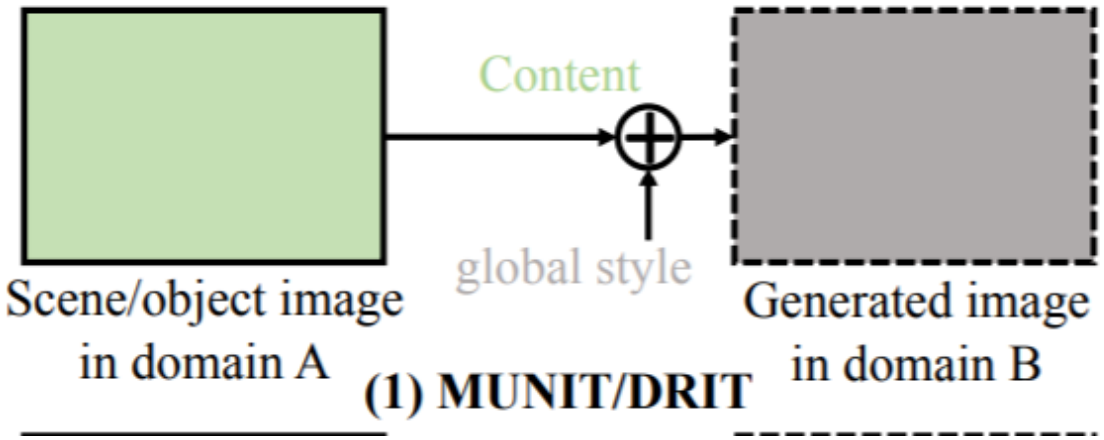
- image sentiment transfer at the object level, leveraging image captioning, semantic image segmentation, and image-to-image translation.
- object level sentiment transfer, SentiGAN (object-level loss, content disentanglement loss)
- create an object-oriented image sentiment dataset to train the image sentiment transfer models.

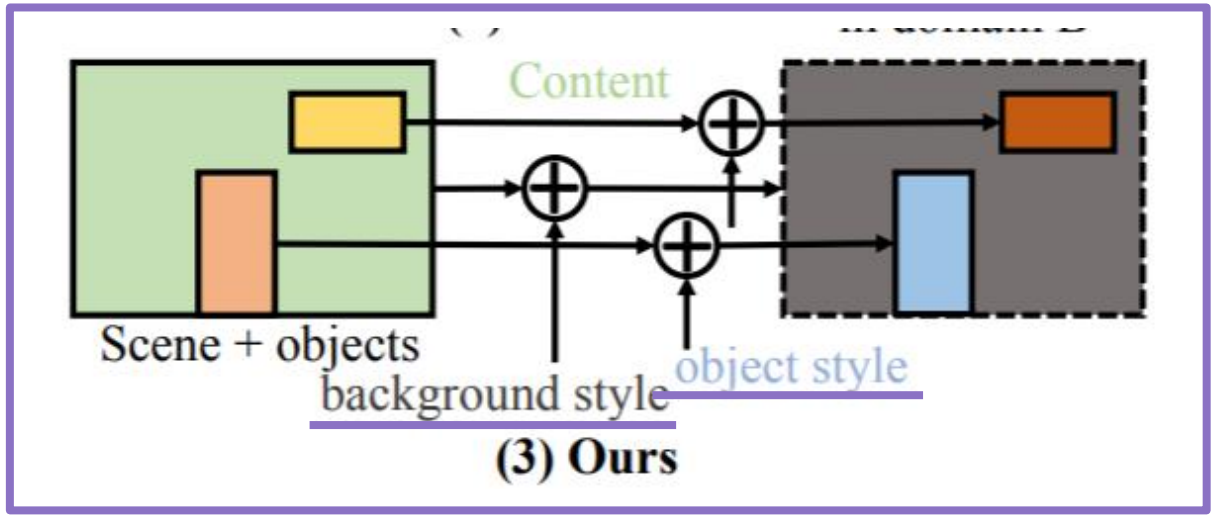


Related Works

1) visual sentiment recognition
   - Low-level (color, texture, shape)
   - Mid-level (composition, sentributes, principles-of-arts)
   - High-level (adjective noun pairs (ANP) e.g., cute dog, beautiful landscape)

2) Towards Instance-level Image-to-Image Translation (INIT) link





→ but dataset is too simple , only street & car

Methodology

Overall Framework

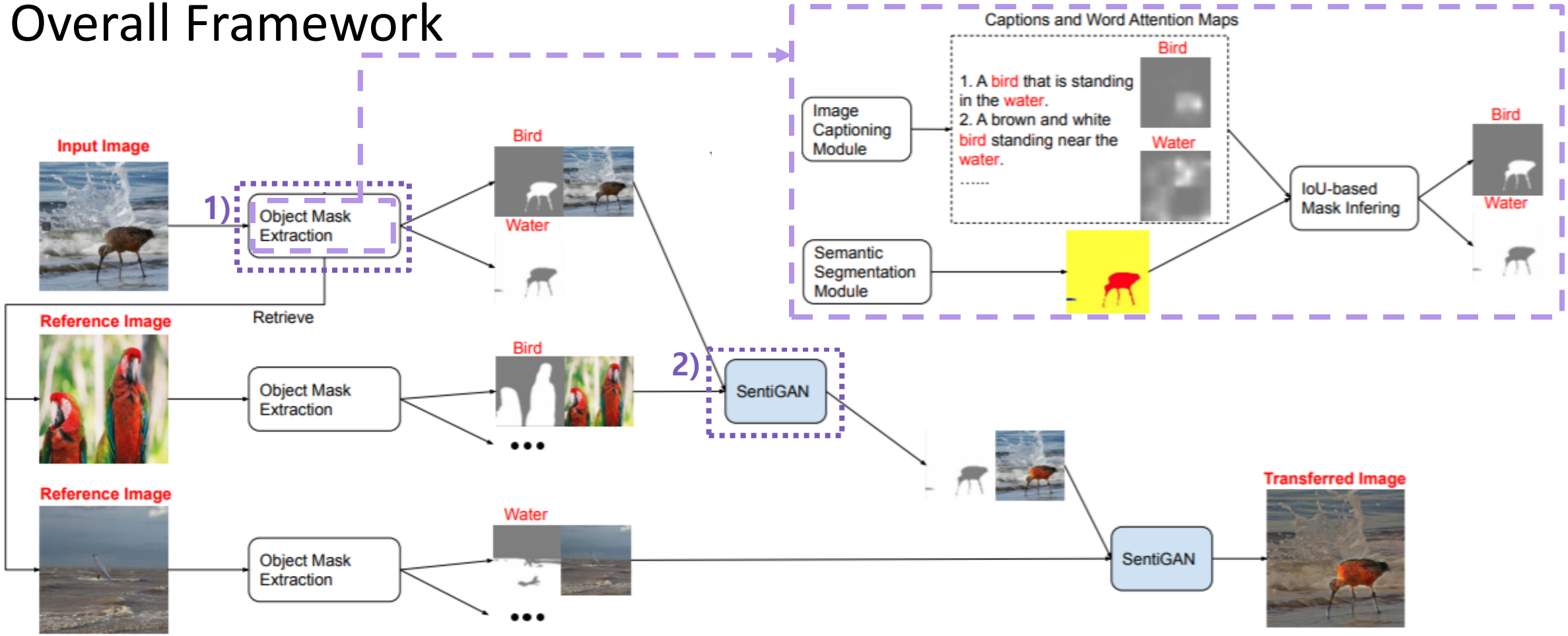
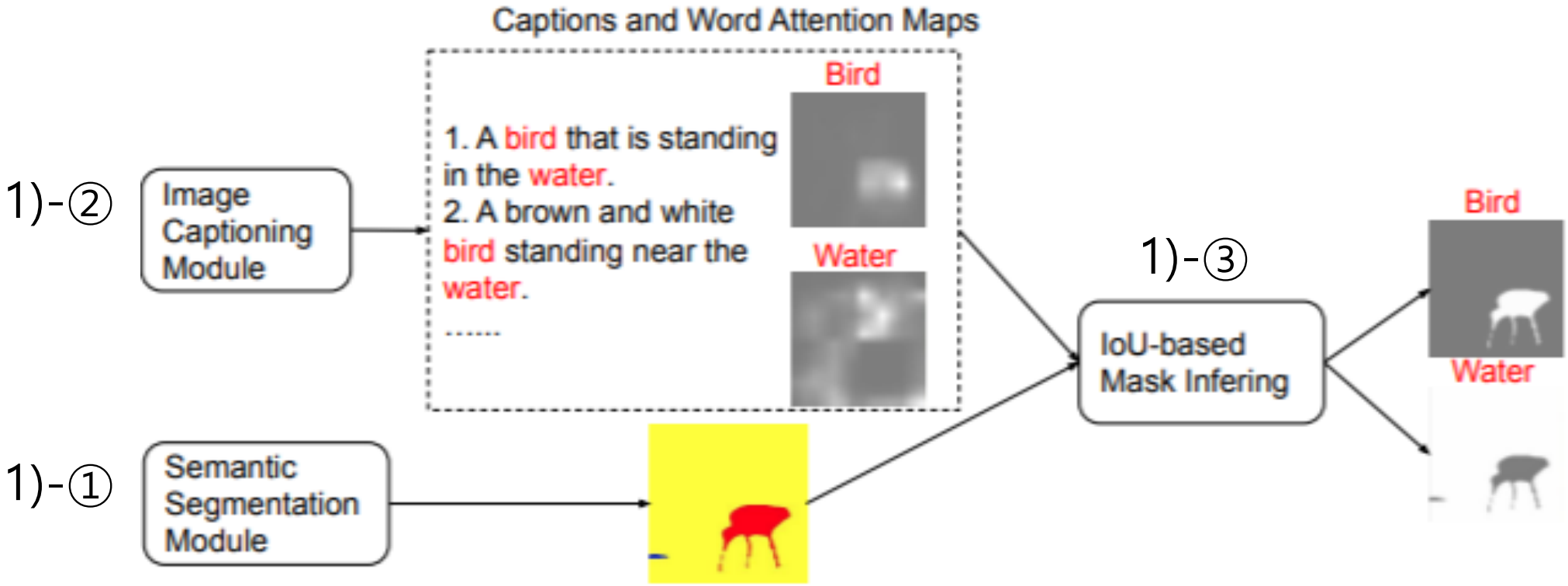


Figure 3: The pipeline of the proposed framework. Given an input image, object mask extraction is first performed to extract the objects and the corresponding masks. Image captioning and semantic image segmentation are utilized to obtain comprehensive objects and high-quality masks. After that, object-level sentiment transfer is performed object-by-object by SentiGAN.

Methodology

1) Object Mask Extraction

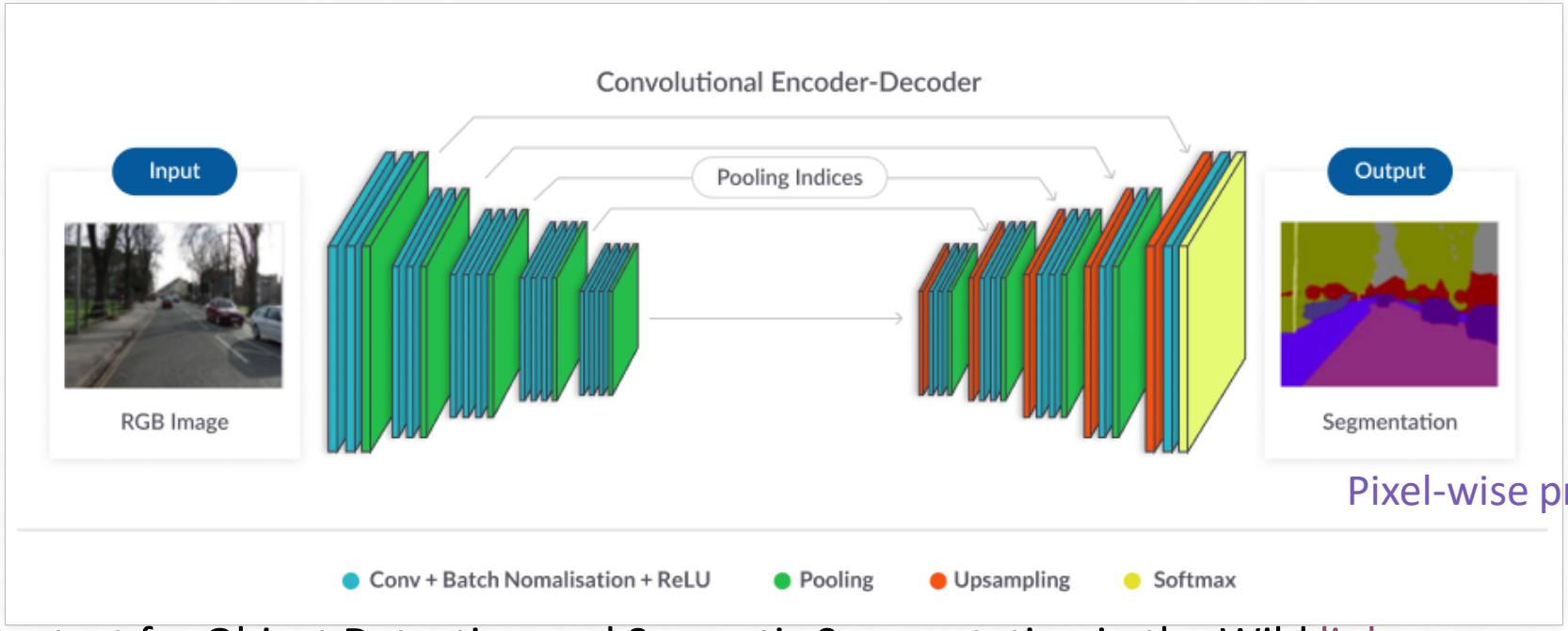


Methodology

1) Object Mask Extraction

1)-① Pre-trained semantic segmentation model

→ **pixel-level** masks of all the present objects



Pixel-wise prediction

- The Role of Context for Object Detection and Semantic Segmentation in the Wild link
- Microsoft COCO: Common Objects in Context link
- Scene Parsing through ADE20K Dataset link

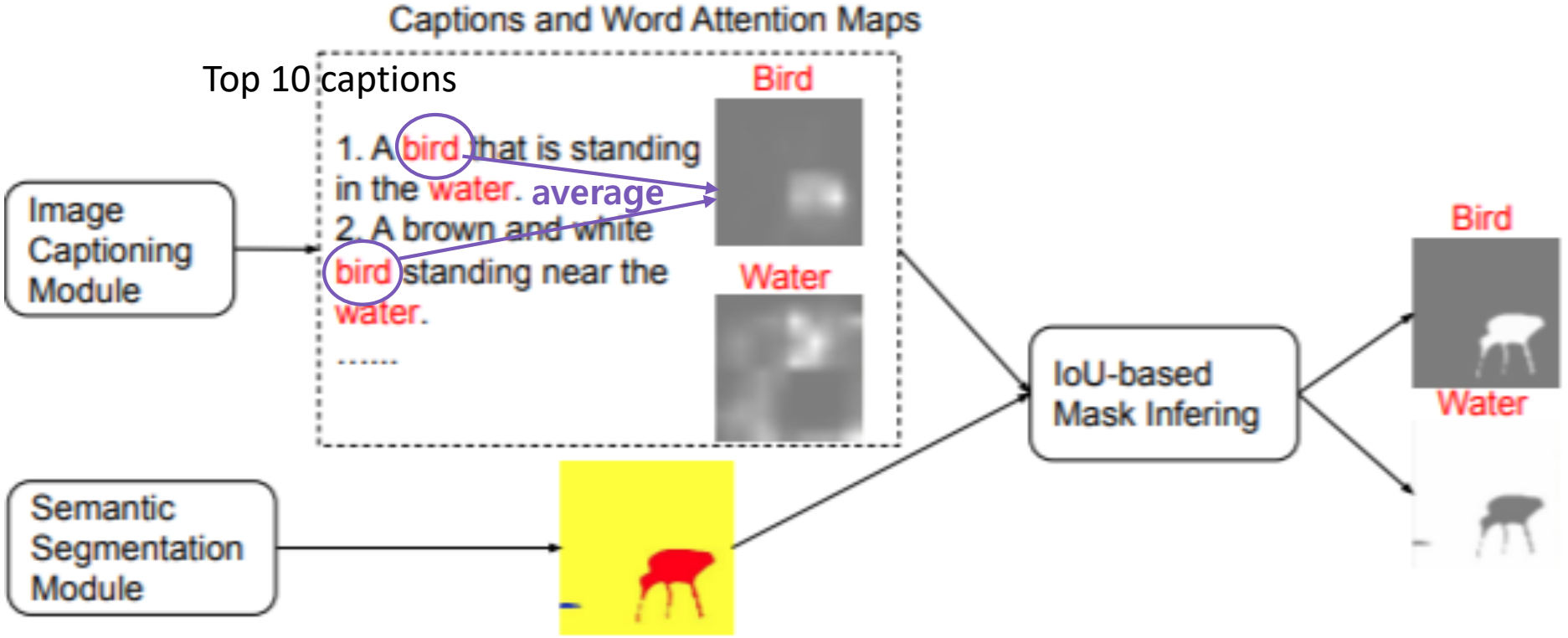
Pre-trained on only a few datasets
→ Limited object classes

Methodology

1) Object Mask Extraction

1)-② Attention-based Image captioning

- Premise
  - semantic segmentation model outputs accurate segmentation for the undefined object based on its learned knowledge on edge detection.



Methodology

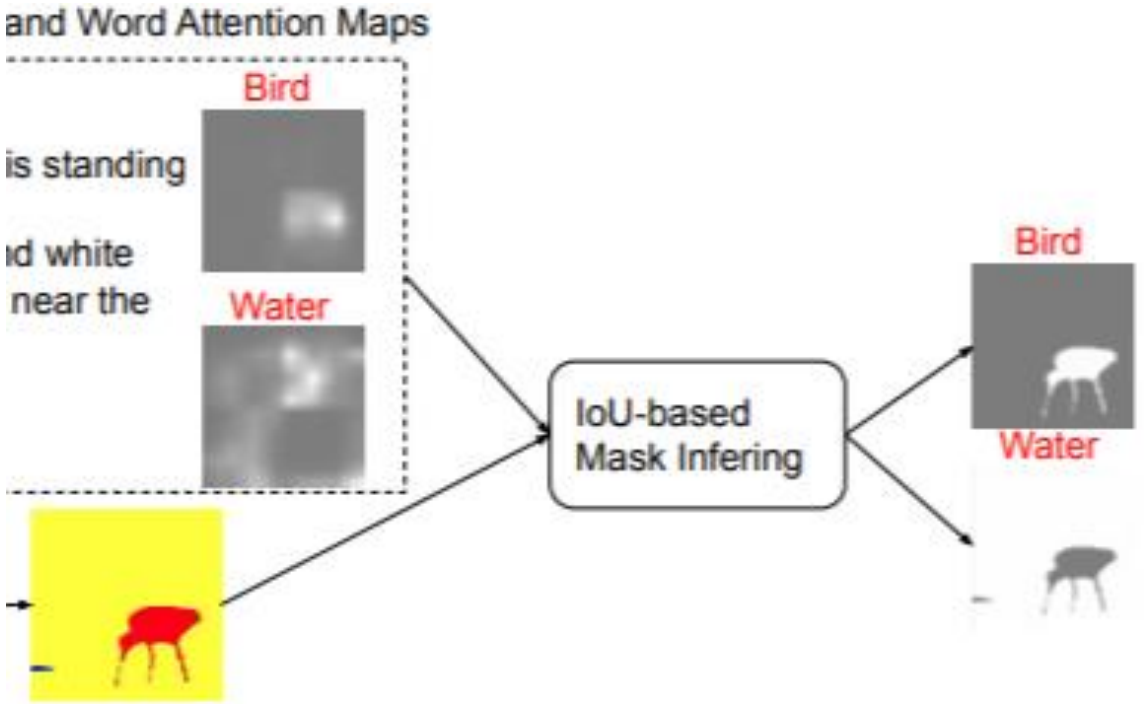
1) Object Mask Extraction

1)-③ IoU-based mask inferring

- Class for each object in S

$$= argmax_{c \in C} \frac{(\sum_{w=0}^{W-1} \sum_{h=0}^{H-1} \mathbb{I}(S_{w,h}=c) A_{w,h})^{\alpha}}{\sum_{w=0}^{W-1} \sum_{h=0}^{H-1} \mathbb{I}(S_{w,h}=c)}$$

  - $\mathbb{I}(x)$: the indicator function
    - $\mathbb{I}(x) = 1$ if x is true, $\mathbb{I}(x) = 0$ o/w
  - $S_{w,h}$ & $A_{w,h}$: the value of point (w,h) of S & A
    - S: Semantic segmentation map
    - A: Attention map
  - C: the segmentation class set
  - α: hyper-parameter
    - if α = 1, the highest average attention values
    - if α is extremely large, the highest sum of attention values

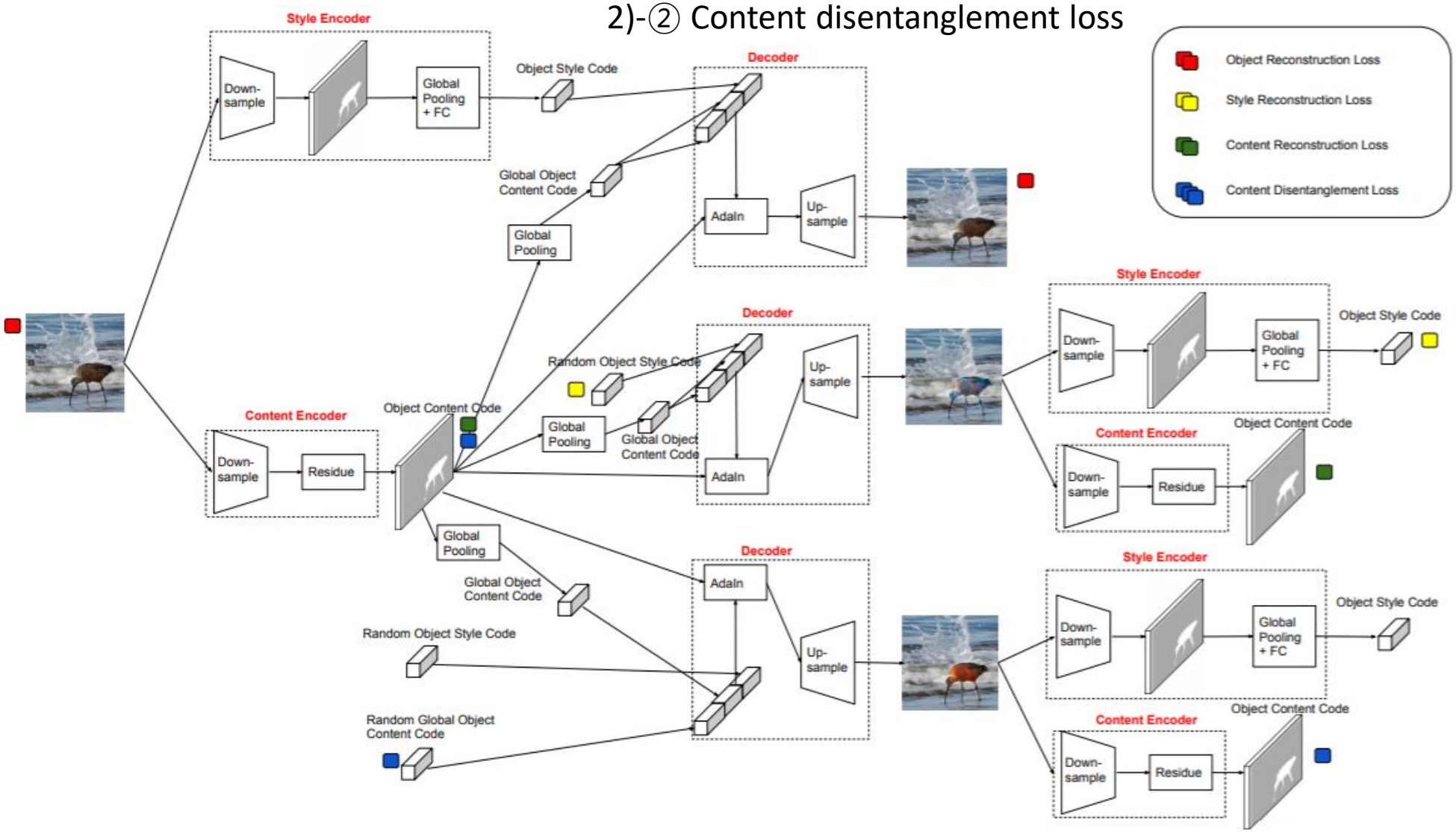


Methodology

2) SentiGAN

2)-① Image-level and Object-level supervision

2)-② Content disentanglement loss



Methodology

2) SentiGAN → based on MUNIT [link]

2)-① Image-level and Object-level supervision

- Image-level

adversarial loss

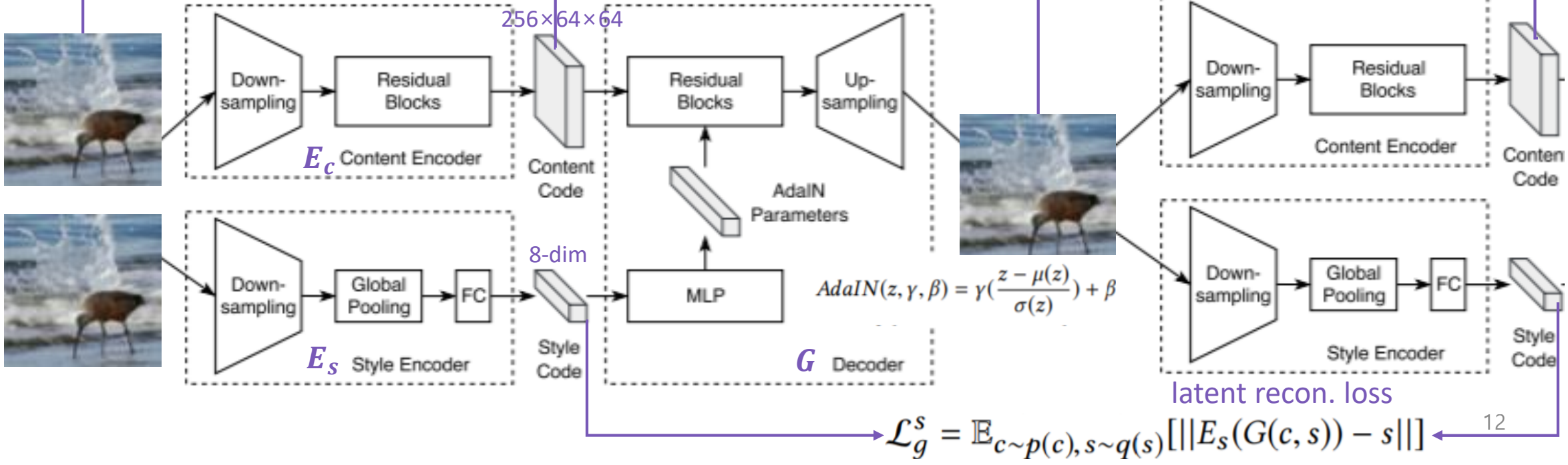
$$+\ \mathcal{L}_g^{gan} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[\log 1 - D(G(c,s))] + \mathbb{E}_{i\sim p(i)}[\log D(i)]$$

Image recon. loss

$$\mathcal{L}_g^{m} = \mathbb{E}_{i\sim p(i)}[||G(E_c(i), E_s(i)) - i||]$$

latent recon. loss

$$\mathcal{L}_g^{c} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||E_c(G(c,s)) - c||]$$



latent recon. loss

$$\mathcal{L}_g^{s} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||E_s(G(c,s)) - s||]$$

Methodology

2) SentiGAN

2)-① Image-level and Object-level supervision

- Image-level

$\mathcal{L}_g^m = \mathbb{E}_{i\sim p(i)}[||G(E_c(i), E_s(i)) - i||]$ —— image reconstruction loss

$i\sim p(i)$: an image sampled from the data distribution

$\mathcal{L}_g^c = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||E_c(G(c,s)) - c||]$

$\mathcal{L}_g^s = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||E_s(G(c,s)) - s||]$ } latent reconstruction loss

$p(c)$: $c = E_c(i), i\sim p(i)$ / $q(s)$: the prior $N(0,I)$

$\mathcal{L}_g^{gan} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[\log 1 - D(G(c,s))] + \mathbb{E}_{i\sim p(i)}[\log D(i)]$ —— adversarial loss

- Object-level

$\boldsymbol{L_o^m} = \mathbb{E}_{i\sim p(i)}[||\boldsymbol{G^o}(E_c(i), \boldsymbol{E_s^o}(i)) - i||]$

$\boldsymbol{L_o^c} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||E_c(\boldsymbol{G^o}(c,s)) - c||]$

$\boldsymbol{L_o^s} = \mathbb{E}_{c\sim p(c), s\sim q(s)}[||\boldsymbol{E_s^o}(\boldsymbol{G^o}(c,s)) - s||]$

E_s & E_s^o shares params

G & G^o shares params

Methodology

2) SentiGAN

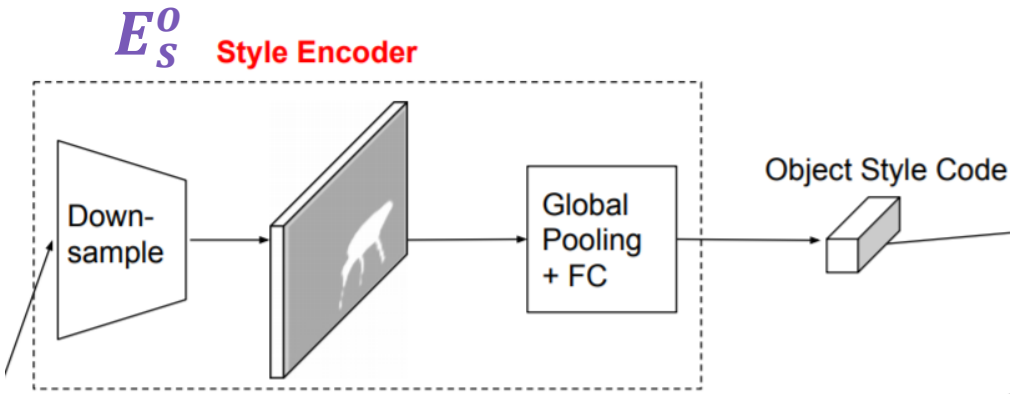
2)-① Image-level and Object-level supervision

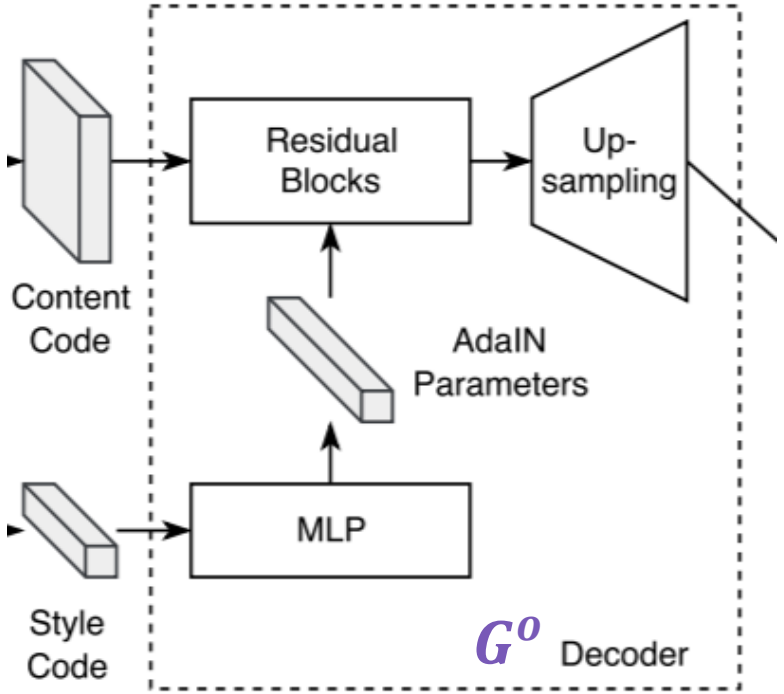
- Object-level

$$L_o^m = \mathbb{E}_{i\sim p(i)}[||G^o(E_c(i), E_s^o(i)) - i||]$$

$$L_o^c = \mathbb{E}_{c\sim p(c),s\sim q(s)}[||E_c(G^o(c,s)) - c||]$$

$$L_o^s = \mathbb{E}_{c\sim p(c),s\sim q(s)}[||E_s^o(G^o(c,s)) - s||]$$





Computed based on the position of the object in z

$$AdaIN(z,\gamma,\beta) = \gamma\left(\frac{z-\mu(z)}{\sigma(z)}\right) + \beta$$

Methodology

2) SentiGAN

2)-② Content disentanglement loss

- Style code does not contain spatial information → no color-based information
- Content code cannot disentangle color-based information

Solution 1: content alignment step (CA) - linear mapping (channel-wise mean & std) → fail
Solution 2: content disentanglement loss

Methodology

2) SentiGAN

2)-② Content disentanglement loss

to transfer color-based information from the referenced image

$$\mathcal{L}_g^{cd} = \mathbb{E}_{c\sim p(c), s\sim q(s), c_{rand}\sim q(c_{rand})}[||\mu(c_{rec}) - \mu(c_{rand})|| + ||\sigma(c_{rec}) - \sigma(c_{rand})|| + ||\frac{c_{rec} - \mu(c_{rec})}{\sigma(c_{rec})} - \frac{c - \mu(c)}{\sigma(c)}||],$$

$$\text{where } c_{rec} = E_c(G(c, s, P(c), P(c_{rand})))$$

P: Global Pooling operation

- Modifying the global spatial-unaware information of the content code does not lead to the loss of the object details.
- The color distributions of the object can be modified by activating specific channels of its content code.

Methodology

2) SentiGAN

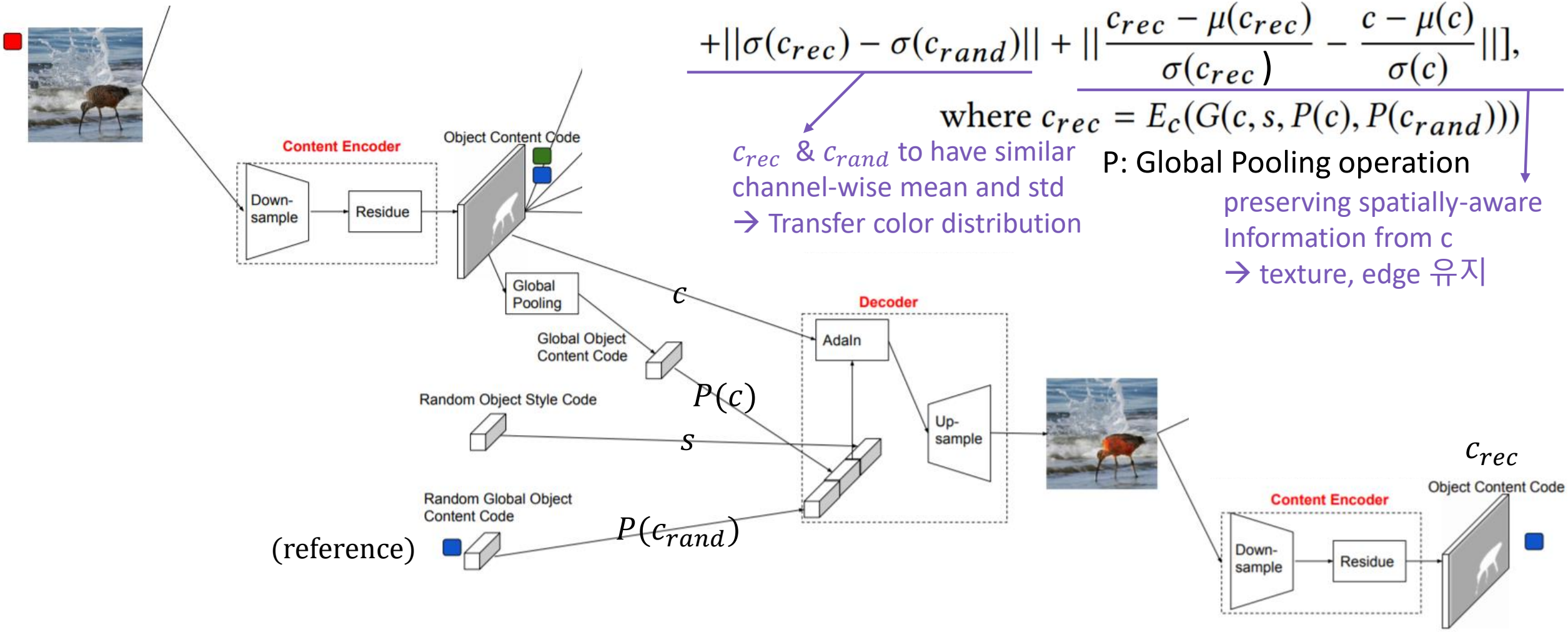
2)-② Content disentanglement loss

$$\mathcal{L}_g^{cd} = \mathbb{E}_{c\sim p(c), s\sim q(s), c_{rand}\sim q(c_{rand})}[||\mu(c_{rec}) - \mu(c_{rand})|| + ||\sigma(c_{rec}) - \sigma(c_{rand})|| + ||\frac{c_{rec} - \mu(c_{rec})}{\sigma(c_{rec})} - \frac{c - \mu(c)}{\sigma(c)}||],$$

where $c_{rec} = E_c(G(c, s, P(c), P(c_{rand})))$

c_{rec} & c_{rand} to have similar channel-wise mean and std
→ Transfer color distribution

P: Global Pooling operation
preserving spatially-aware Information from c
→ texture, edge 유지



Methodology

Overall Framework

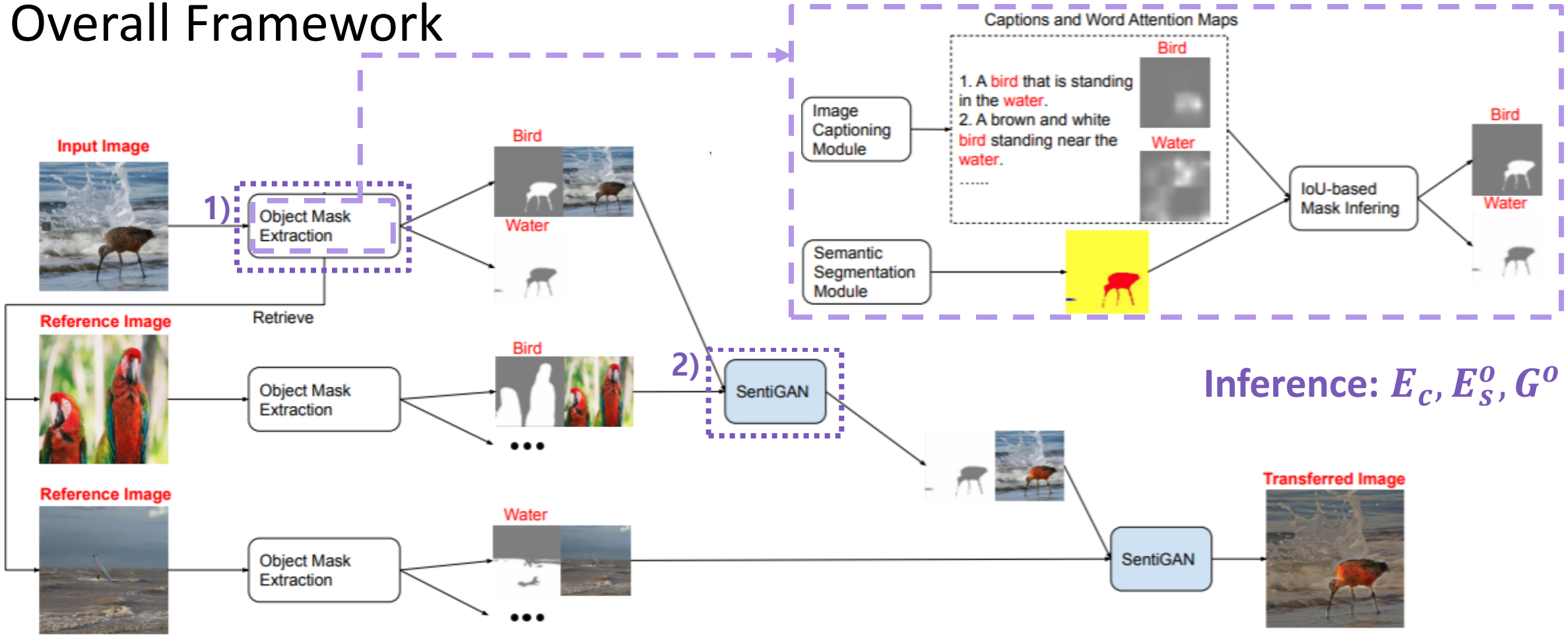


Figure 3: The pipeline of the proposed framework. Given an input image, object mask extraction is first performed to extract the objects and the corresponding masks. Image captioning and semantic image segmentation are utilized to obtain comprehensive objects and high-quality masks. After that, object-level sentiment transfer is performed object-by-object by SentiGAN.

Experiment & Results

Dataset: Visual Sentiment Ontology (VSO)

= Flickr images queried by adjective noun pairs (ANP)

Task 1) coarse-level sentiment transfer

images with neutral sentiment → positive/negative image

Input Images	Positive Rate 0.540	Negative Rate 0.460	
	True Positive Rate	True Negative Rate	Average
MUNIT	0.582	0.478	0.530
MUNIT + ObjSup	0.578	0.484	0.531
MUNIT + ObjSup + CA	0.622	0.484	0.553
SentiGAN - CA	0.594	0.502	0.548
SentiGAN (IDL)	0.580	0.506	0.543
SentiGAN	0.596	0.520	**0.558**

content alignment step (CA) - linear mapping (channel-wise mean & std)

Binary classification

(P or N)

	Hit Rate	Miss Rate
User Study	0.724	0.276

User study

select P in P-N pair

Experiment & Results

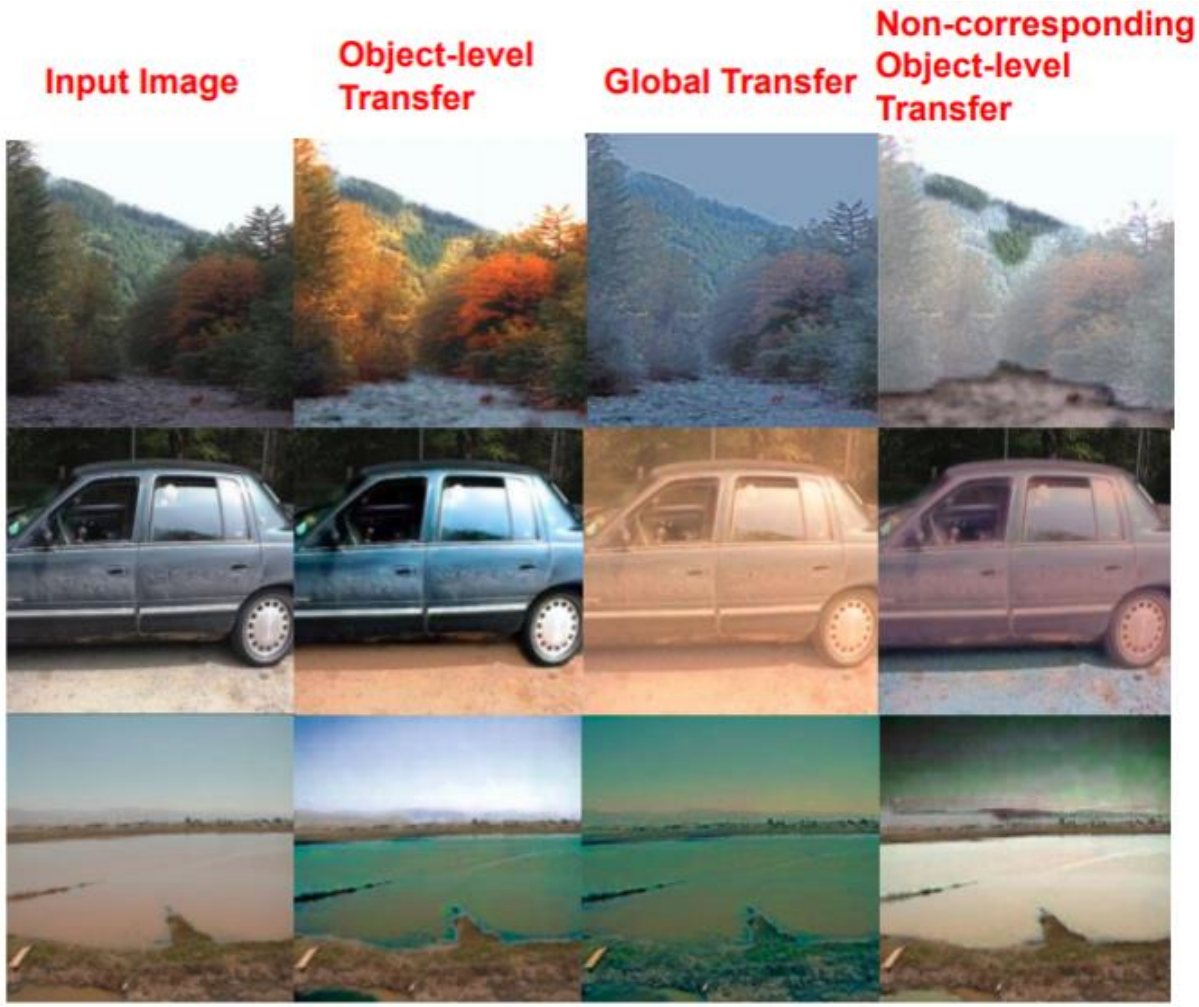
Task 2) effectiveness of the object-level transfer

	Object-level Transfer	Global Transfer
User Study	0.672	0.288

Non-corresponding Object-level Transfer
0.040

User study

Select the most real image



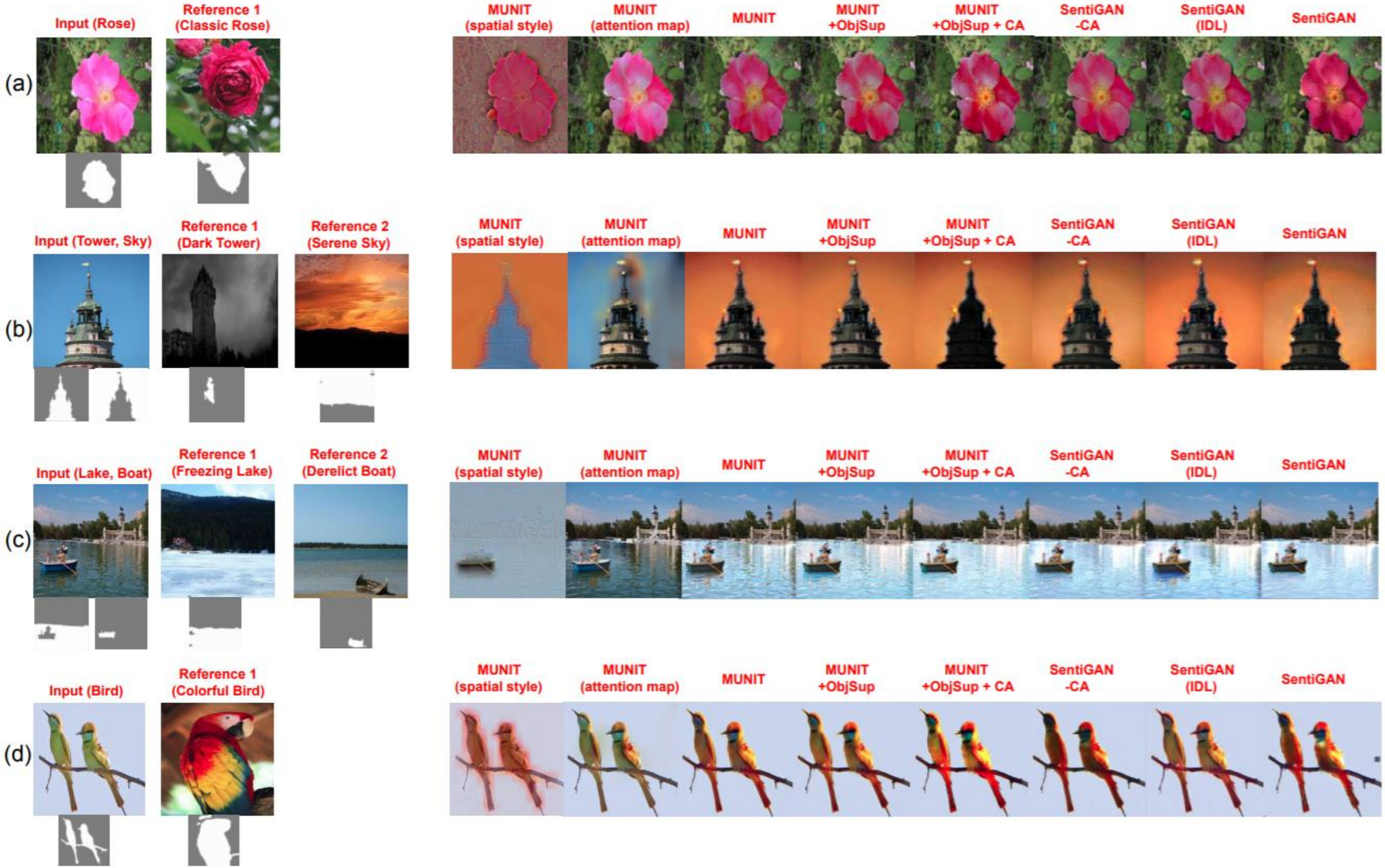
Experiment & Results

Task 3) sentiment consistency between transferred & reference

	Hit Rate
MUNIT	0.129
MUNIT + ObjSup	0.150
MUNIT + ObjSup + CA	0.189
SentiGAN - CA	0.184
SentiGAN (IDL)	0.123
SentiGAN	**0.226**

User study

select the most consistent
with reference


(a)
Input (Rose)
Reference 1 (Classic Rose)
MUNIT (spatial style)
MUNIT (attention map)
MUNIT
MUNIT +ObjSup
MUNIT +ObjSup + CA
SentiGAN -CA
SentiGAN (IDL)
SentiGAN
(b)
Input (Tower, Sky)
Reference 1 (Dark Tower)
Reference 2 (Serene Sky)
MUNIT (spatial style)
MUNIT (attention map)
MUNIT
MUNIT +ObjSup
MUNIT +ObjSup + CA
SentiGAN -CA
SentiGAN (IDL)
SentiGAN
(c)
Input (Lake, Boat)
Reference 1 (Freezing Lake)
Reference 2 (Derelict Boat)
MUNIT (spatial style)
MUNIT (attention map)
MUNIT
MUNIT +ObjSup
MUNIT +ObjSup + CA
SentiGAN -CA
SentiGAN (IDL)
SentiGAN
(d)
Input (Bird)
Reference 1 (Colorful Bird)
MUNIT (spatial style)
MUNIT (attention map)
MUNIT
MUNIT +ObjSup
MUNIT +ObjSup + CA
SentiGAN -CA
SentiGAN (IDL)
SentiGAN

참고 자료

Category	Short Name	#	Short Description
color	*Saturation, Brightness*	2	mean saturation and brightness
	Pleasure, Arousal, Dominance	3	approx. emotional coordinates based on brightness and saturation
	Hue	4	vector based mean hue and angular dispersion, saturation weighted and without saturation
	Colorfulness	1	colorfulness measure based on EMD
	Color Names	11	amount of black, blue, brown, green, gray, orange, pink, purple, red, white, yellow
	Itten	20	average *contrast of brightness, contrast of saturation, contrast of hue, contrast of complements, contrast of warmth, harmony, hue count, hue spread, area of warm, area of cold,*...and the maximum of each
	Wang	19	features (histograms) by Wang Wei-ning et al. [31] (*factors 1 (10), factor 2 (7) and factor 3 (2)*)
	Area statistics	10	based on Wang features: *area of very dark, area of dark, area of middle, area of...light, very light, high saturation, middle saturation, low saturation, warm, cold*
texture	*Tamura*	3	features by Tamura et al [25]: *coarseness, contrast, directionality*
	Wavelet textures	12	wavelet textures for each channel (Hue, Saturation, Brightness) and each level (1-3), sum of all levels for each channel
	GLCM-features	12	features based on the GLCM: *contrast, correlation, energy, homogeneity* for Hue, Saturation and Brightness channel
composition	*Level of Detail*	1	number of segments after waterfall segmentation
	Low Depth of Field (DOF)	3	low depth of field indicator; ratio of wavelet coefficients of inner rectangle vs. whole image (for Hue, Saturation and Brightness channel)
	Dynamics	6	Line slopes: static, dynamic (absolute and relative), lengths of static lines, lengths of dynamic lines
	Rule of Thirds	3	mean saturation, brightness and hue of the inner rectangle
content	*Faces*	2	number of frontal faces, relative size of the biggest face
	Skin	2	number of skin pixels, relative amount of skin with respect to the size of faces

Table 1: Summary of all features. The column '#' indicates the feature vector length for each type of feature.

Affective Image Classification using Features Inspired by Psychology and Art Theory

Jana Machajdik, Allan Hanbury link

- **Sentributes**

  (1) Material: such as metal, vegetation
  (2) Function: playing, cooking
  (3) Surface property: rusty, glossy
  (4) Spatial Envelope: natural, man-made, enclosed.

Sentribute: Image Sentiment Analysis from a Mid-level Perspective
Jianbo Yuan et al. link

- **principles-of-arts**

Table 2: Summary of the measurements for principles of art. '#' indicates the dimension of each measurement.

Principles	Measurement	#	Short Description
Balance	*Bilateral symmetry*	12	Symmetry number, Maximum symmetry radius, angle and strength
	Rotational symmetry	12	Symmetry number, Maximum symmetry center (x and y), strength
	Radial symmetry	36	Distribution of symmetry map after radial symmetry transformation
Emphasis	*Itten color contrast*	15	Average contrast of saturation, contrast of light and dark, contrast of extension, contrast of complements, contrast of hue, contrast of warm and cold, simultaneous contrast
	RFA	3	Rate of focused attention based on saliency map and subject mask
Harmony	*Rangeability of hue and gradient direction*	2	The first and second maximums of local maximum hues and gradient directions in relative histograms of an image patch, and their differences; the combination of all patches of an image
Variety	*Color names*	12	Color types of black, blue, brown, gray, green, orange, pink, purple, red, white, yellow and each color's amount
	Distribution of gradient	48	The distribution of gradient on eight scales of direction and eight scales of length
Gradation	*Absolute and relative variation*	9	Pixel-wise windowed total variation, windowed inherent variation in x and y direction respectively, and relative total variation
Movement	*Gaze scan path*	16	The distribution of gaze vector

Exploring Principles-of-Art Features For Image Emotion Recognition link
Sicheng Zhao† , Yue Gao‡ , Xiaolei Jiang† , Hongxun Yao† , Tat-Seng Chua‡ , Xiaoshuai Sun†

Towards Instance-level Image-to-Image Translation (INIT) link

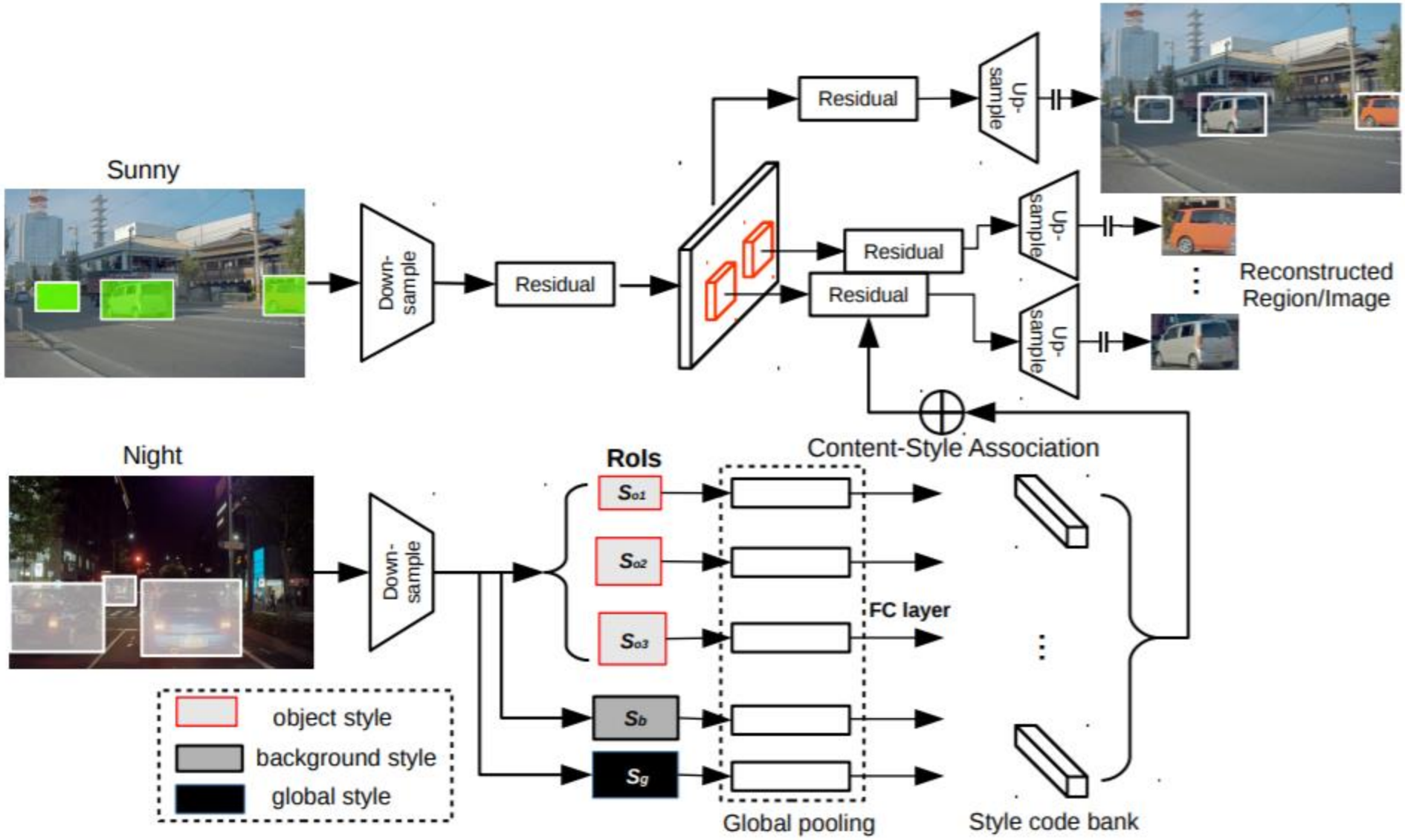


Figure 4. Overview of our instance-aware cross-domain I2I translation. The whole framework is based on the MUNIT method [12], while we further extend it to realize the instance-level translation purpose. Note that after content-style association, the generated images will place in the target domain, so a translation back process will be employed before self-reconstruction, which is not illustrated here.