MMML（Multimodal Machine Learning）

姓名：汪志伟

12.17

Text2Image

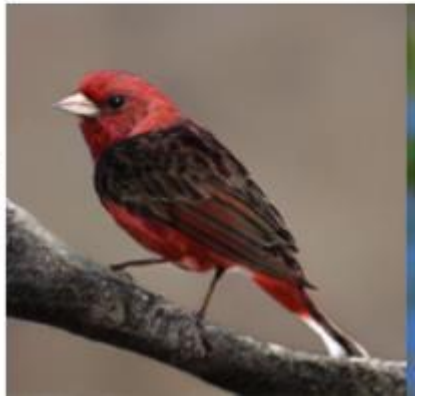
1.Image synthesis

2.text-guided manipulation

3.text-guided inpainting

Image synthesis

There is a red bird with black beady eyes and dark edged feather sitting on a branch.



Text-guided manipulation

A bird with **black eye rings** and a **black bill**, with a **red crown** and a **red belly**.

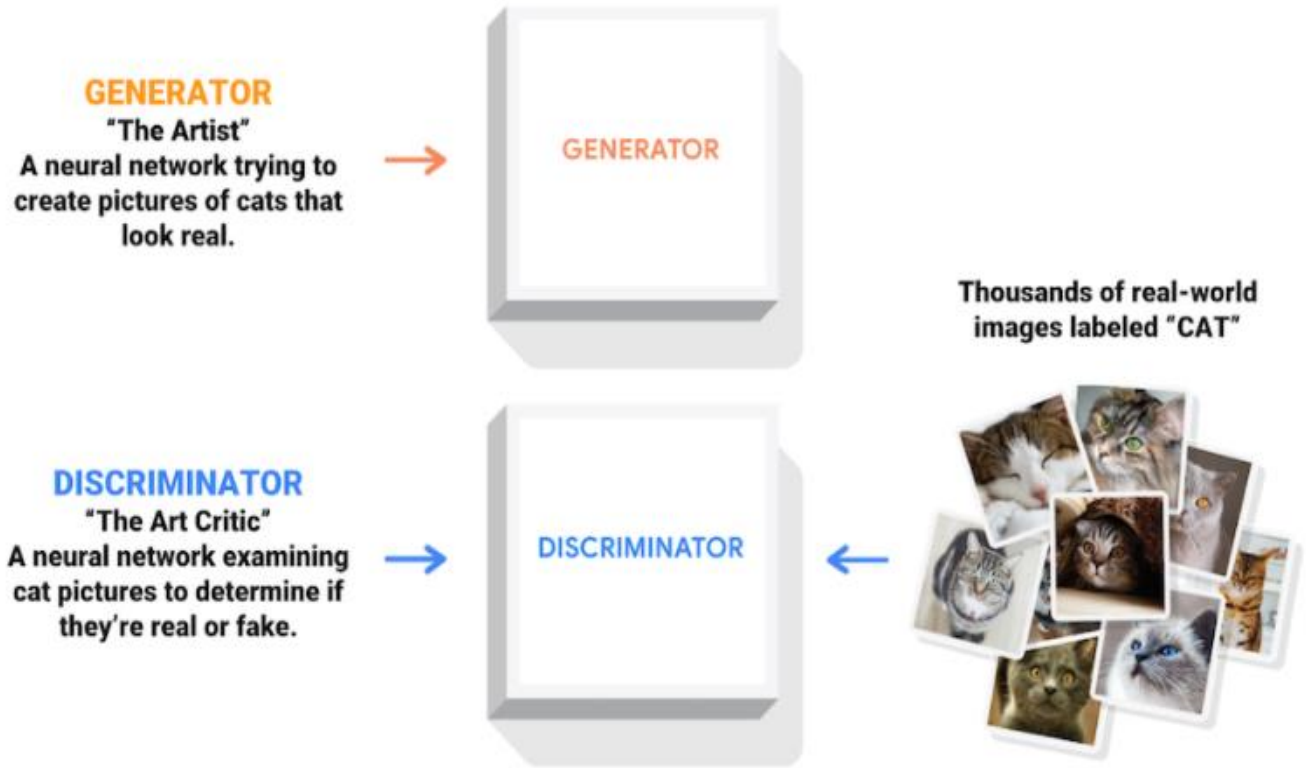
Text-guided inpainting

This is a bird with a white belly, grey and yellow wings and a white eye ring

Elementary Knowledge

- GAN
- Affine Transformation
- Patch_based
- Diffusion_based
- Sentence interpolation
- Dilated convolution

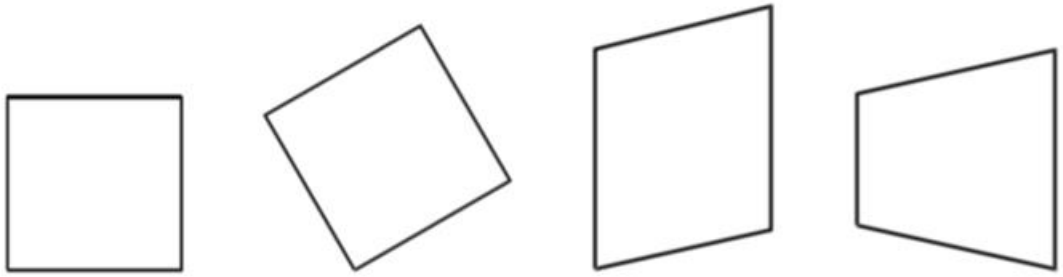
GAN



判别器损失：
该方法量化判别器从判断真伪图片的能力。它将判别器对真实图片的预测值与值全为 1 的数组进行对比，将判别器对伪造（生成的）图片的预测值与值全为 0 的数组进行对比。

生成器损失：
生成器损失量化其欺骗判别器的能力。直观来讲，如果生成器表现良好，判别器将会把伪造图片判断为真实图片（或 1）。这里我们将把判别器在生成图片上的判断结果与一个值全为 1 的数组进行对比。

Affine Transformation



仿射变换（Affine Transformation）
其实是另外两种简单变换的叠加：
一个是线性变换，一个是平移变换

仿射变换中集合中的一些性质保持不变：
（1）凸性
（2）共线性：若几个点变换前在一条线上，则仿射变换后仍然在一条线上
（3）平行性：若两条线变换前平行，则变换后仍然平行
（4）共线比例不变性：变换前一条线上两条线段的比例，在变换后比例不变

作用：这种方法经常在conditional normalization上，以包含一些附加的信息、或者避免因为normalization造成信息损失

Patch_based

将整个图片分成多个图片块，再进行处理。

如：

- 为了自动识别癌症的亚型，在千亿像素级的WSI（Whole Slide Tissue Images）上训练一个CNN是不可能的。癌症亚型的差异是基于细胞级的视觉特征，在图像patch上被观察

- 有些图片虽然标签不同，但局部地区有很多相似的，可以通过对比这些局部相似性和差异性进行学习，获取更多信息。

Diffusion_based

- 利用待修复区域的边缘信息，确定扩散的方向。
- 向边缘内扩散已知的信息。

缺点：在修复区域较大的任务中效果不行，会存在纹理不一致，内容不合理。

Sentence interpolation

Interpolation：

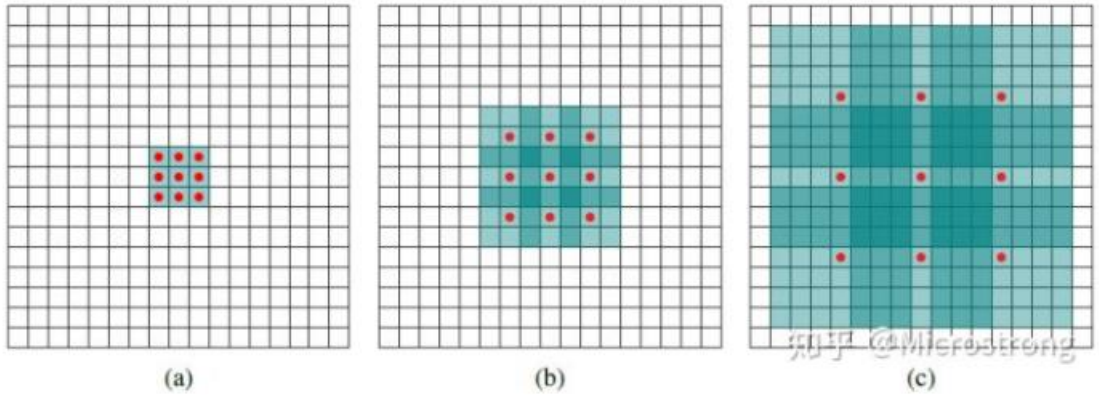
是一种通过已知的、离散的数据点，在范围内推求新数据点的过程或方法。

Type：

1. 片段插值
2. 线性插值
3. 多项式插值
4. 样条曲线插值
5. ……

Dilated convolution

- 标准的卷积操作中，卷积核的元素之间都是相邻的。但是，在空洞卷积中，卷积核的元素是间隔的，间隔的大小取决于空洞率。



ManiGAN: Text-Guided Image Manipulation

Bowen Li[1] Xiaojuan Qi[1,2] Thomas Lukasiewicz[1] Philip H. S. Torr[1]
[1]University of Oxford [2]University of Hong Kong

CVPR 2020

Motivation

1. current state-of-the-art text-guided image manipulation methods are only able to produce low-quality images

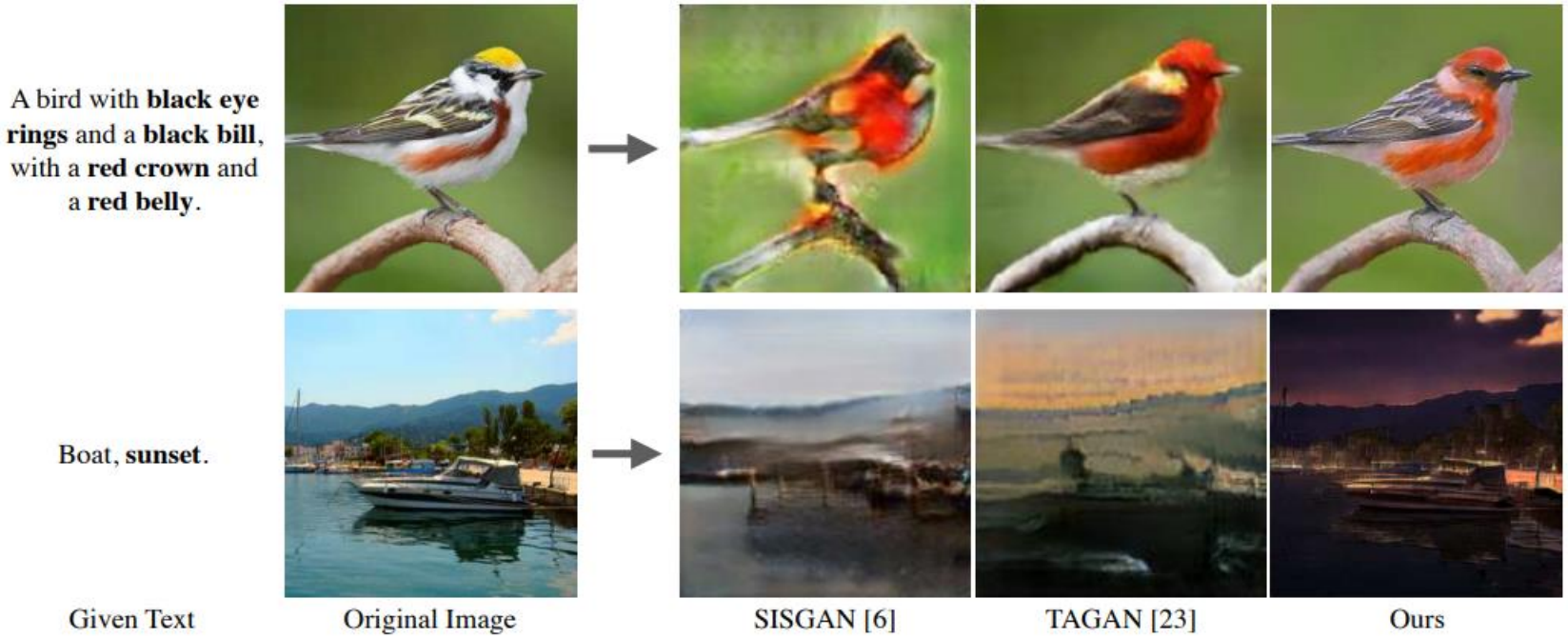
2. cannot precisely correlate fine-grained words with corresponding visual attributes that need to be modified

3. cannot effectively identify text-irrelevant contents and thus fails to reconstruct them

The goal

1. **the task aims to semantically edit parts of an image according to the given text provided by a user**

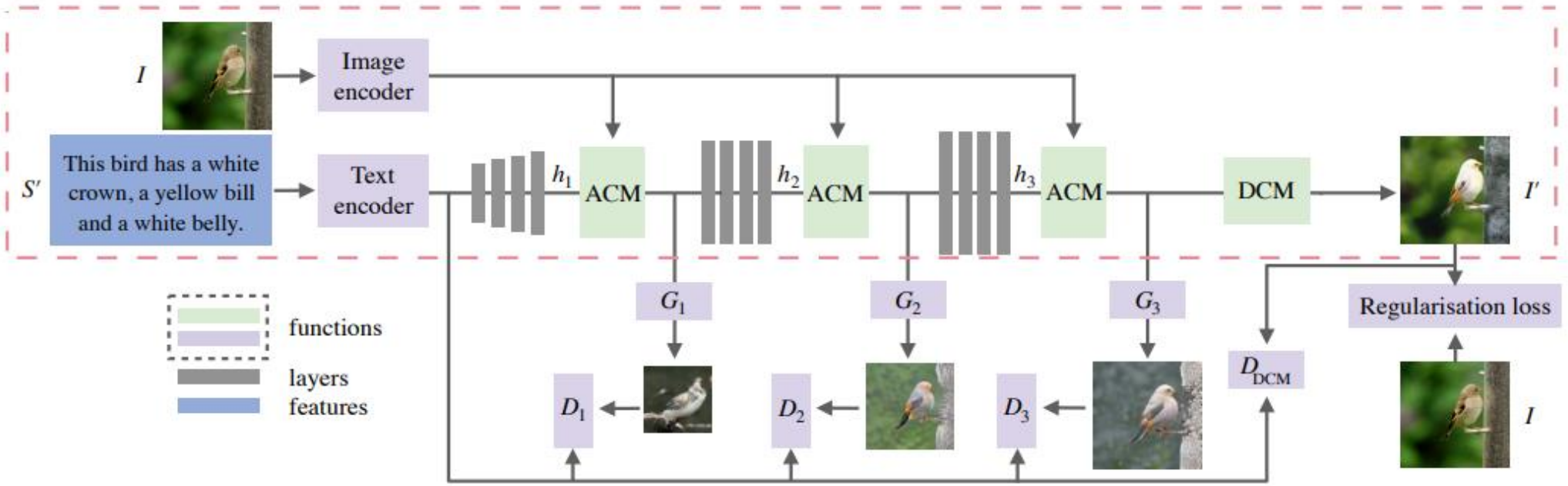
2. **preserve other contents that are not described in the text**



Application

- Video games
- Image editing
- Computer-aided design

Architecture

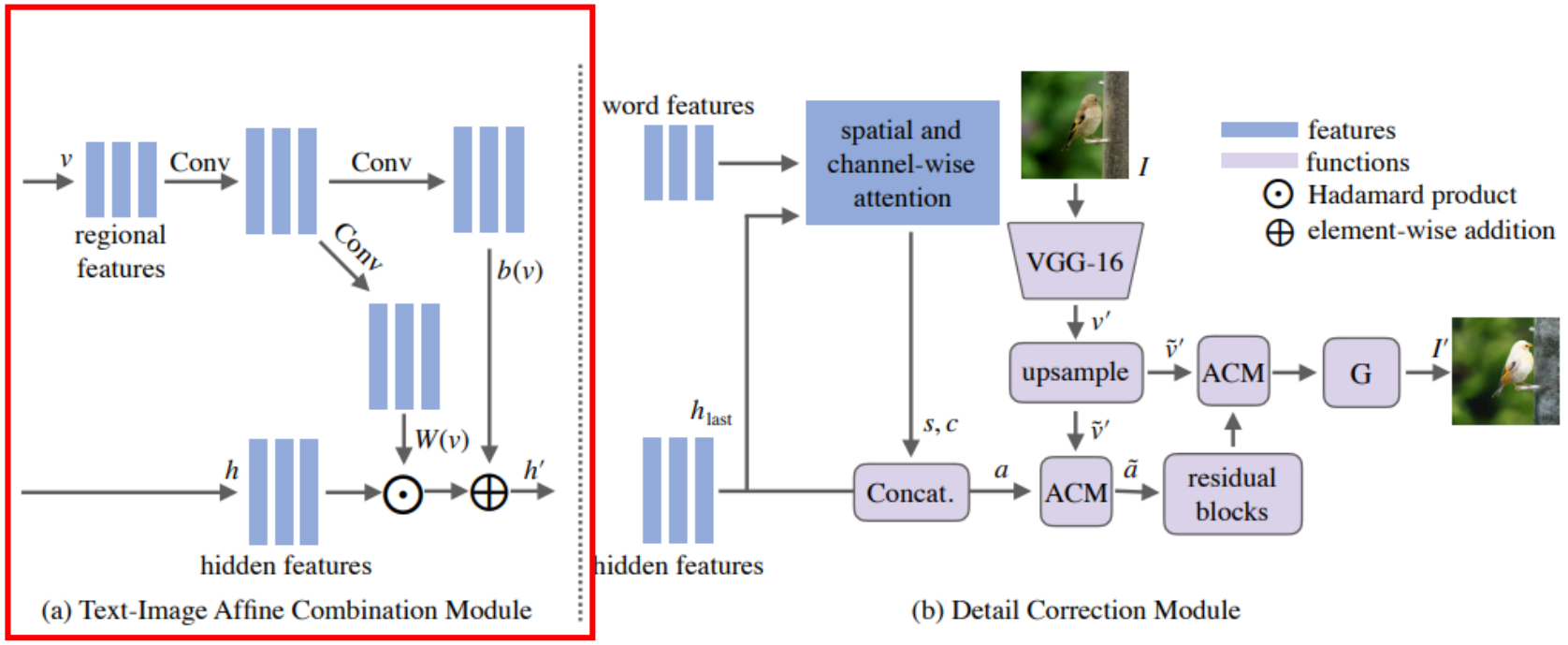


ACM:text-image affine combination module

DCM:detail correction module

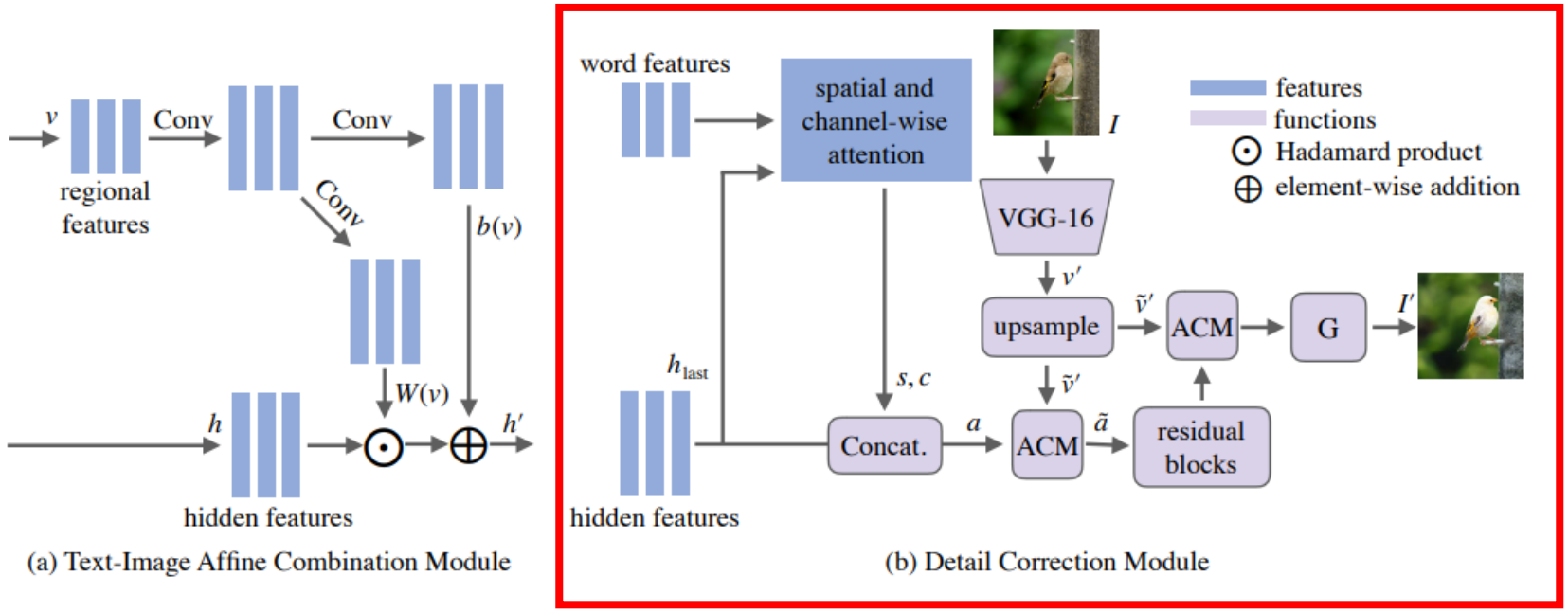
How?

ACM:text-image affine combination module



$$h' = h \odot W(v) + b(v), \qquad (1)$$

DCM: detail correction module



Dataset

Dataset	Training set	Test set	Captions per image
CUB	8855	2933	10
COCO	82785	40504	5

Experiment

Method	CUB				COCO			
	IS	sim	diff	MP	IS	sim	diff	MP
SISGAN [6]	2.24	.045	.508	.022	3.44	.077	.442	.042
TAGAN [23]	3.32	.048	.267	.035	3.28	.089	.545	.040
Ours w/o ACM	4.01	**.138**	.491	.070	5.26	.121	.537	.056
Ours w/ Concat.	3.81	.135	.512	.065	13.48	.085	.532	.039
Ours w/o main	**8.48**	.084	**.235**	.064	**17.59**	.080	**.169**	.066
Ours w/o DCM	3.84	.123	.447	.068	6.99	**.138**	.517	.066
Ours	8.47	.101	.281	**.072**	14.96	.087	.216	**.068**

MP: manipulative precision

$$\mathrm{MP} = (1 - \mathrm{diff}) \times \mathrm{sim}, \tag{3}$$

Experiment

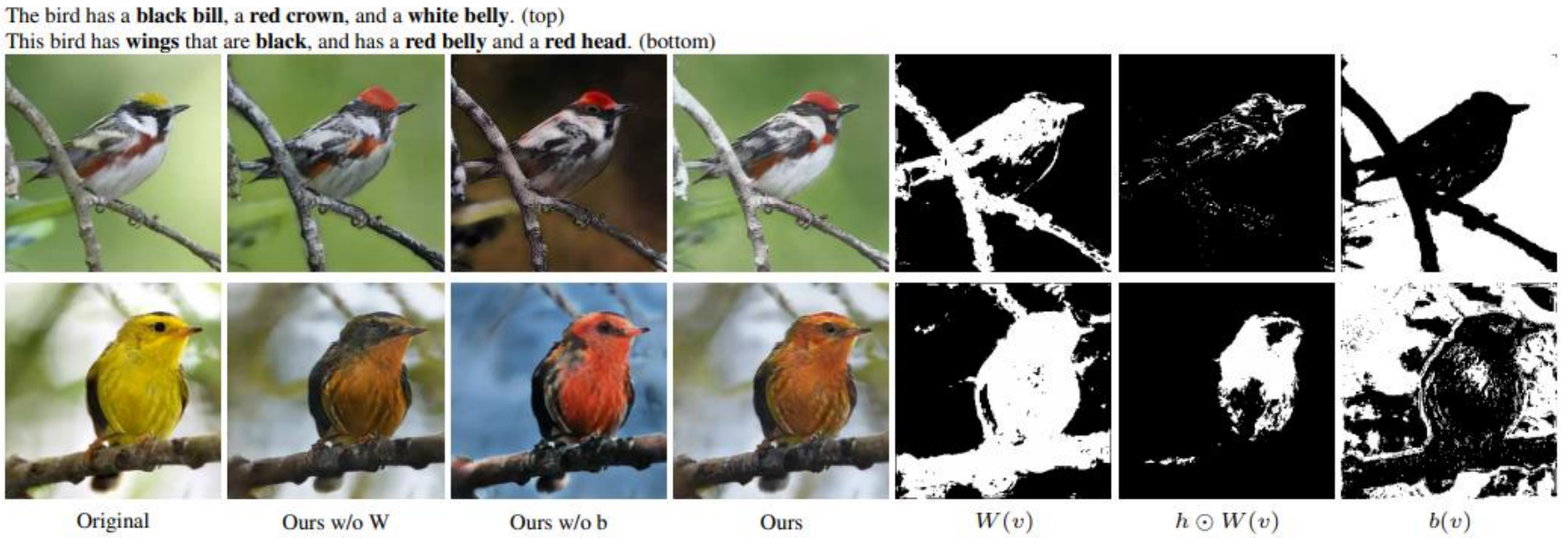
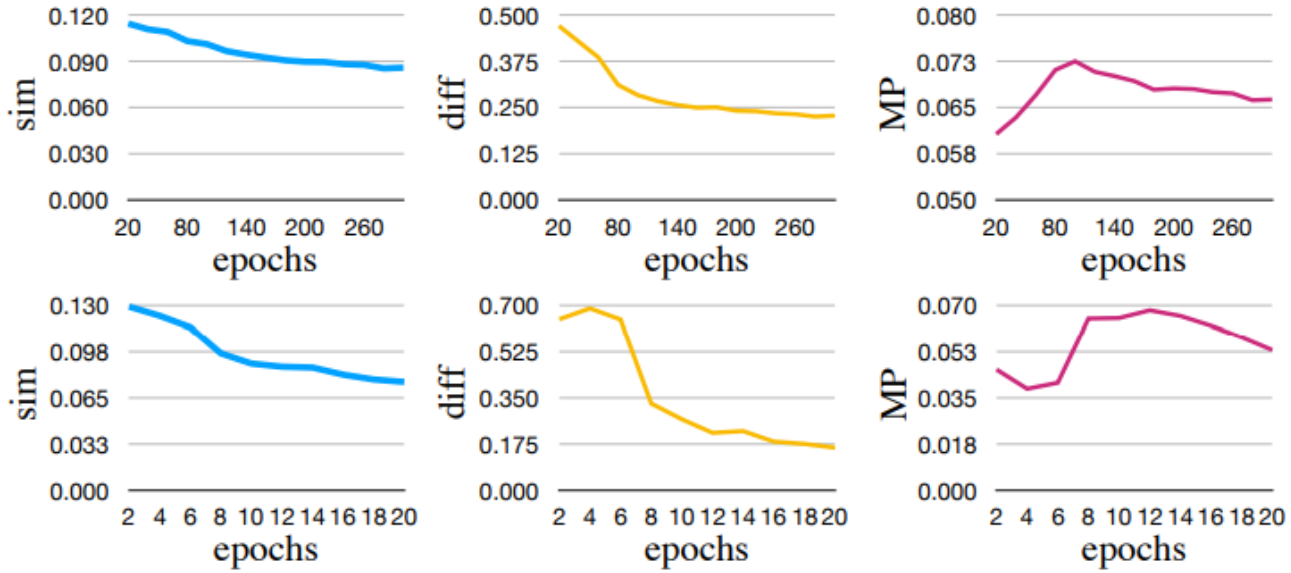


Figure 4: Ablation studies of the learned W and b. The texts on the top are the given descriptions containing desired visual attributes, and the last three columns are the channel feature maps of $W(v)$, $h \odot W(v)$, and $b(v)$.

Experiment


sim
0.120
0.090
0.060
0.030
0.000
20 80 140 200 260
epochs
diff
0.500
0.375
0.250
0.125
0.000
20 80 140 200 260
epochs
MP
0.080
0.073
0.065
0.058
0.050
20 80 140 200 260
epochs
sim
0.130
0.098
0.065
0.033
0.000
2 4 6 8 10 12 14 16 18 20
epochs
diff
0.700
0.525
0.350
0.175
0.000
2 4 6 8 10 12 14 16 18 20
epochs
MP
0.070
0.053
0.035
0.018
0.000
2 4 6 8 10 12 14 16 18 20
epochs

Experiment

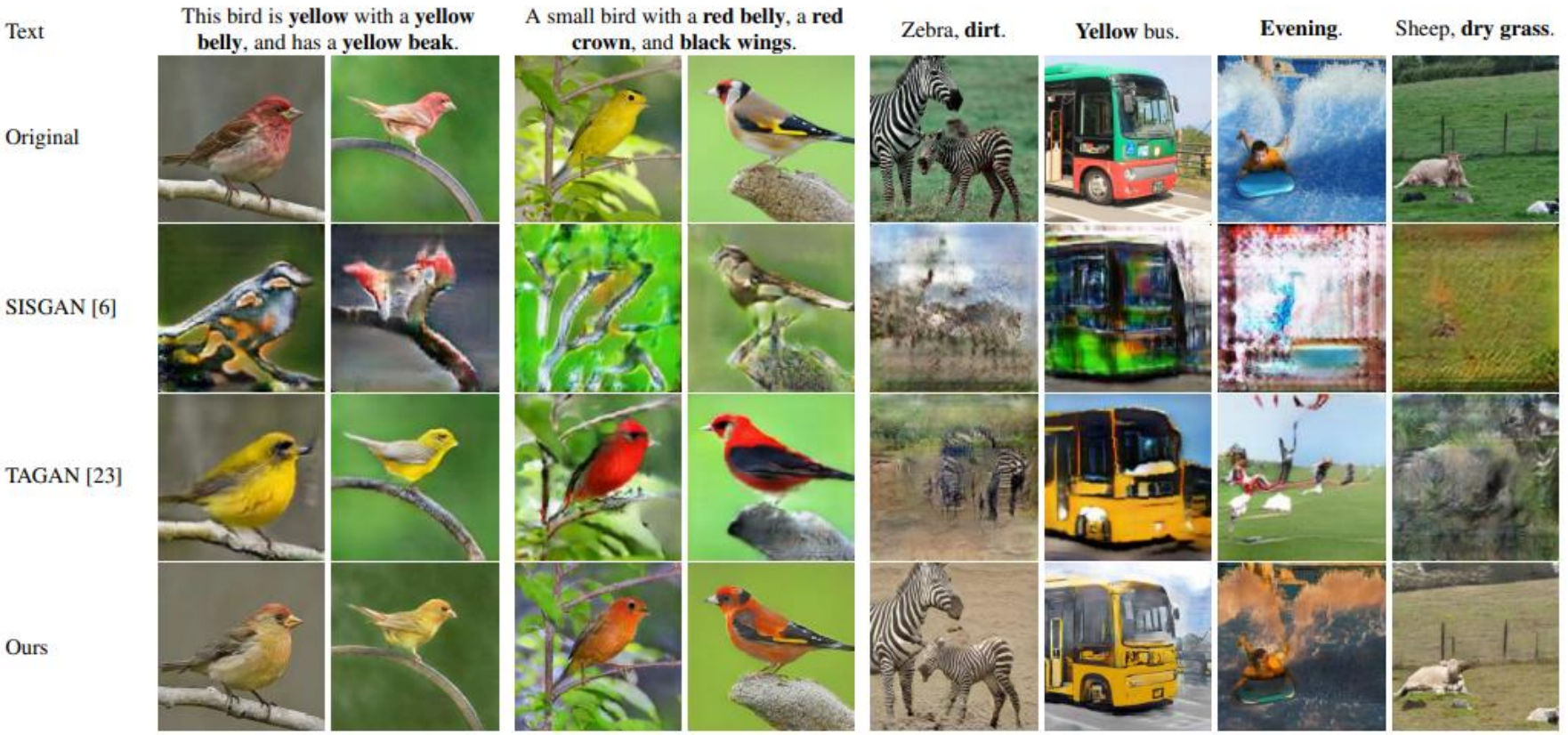
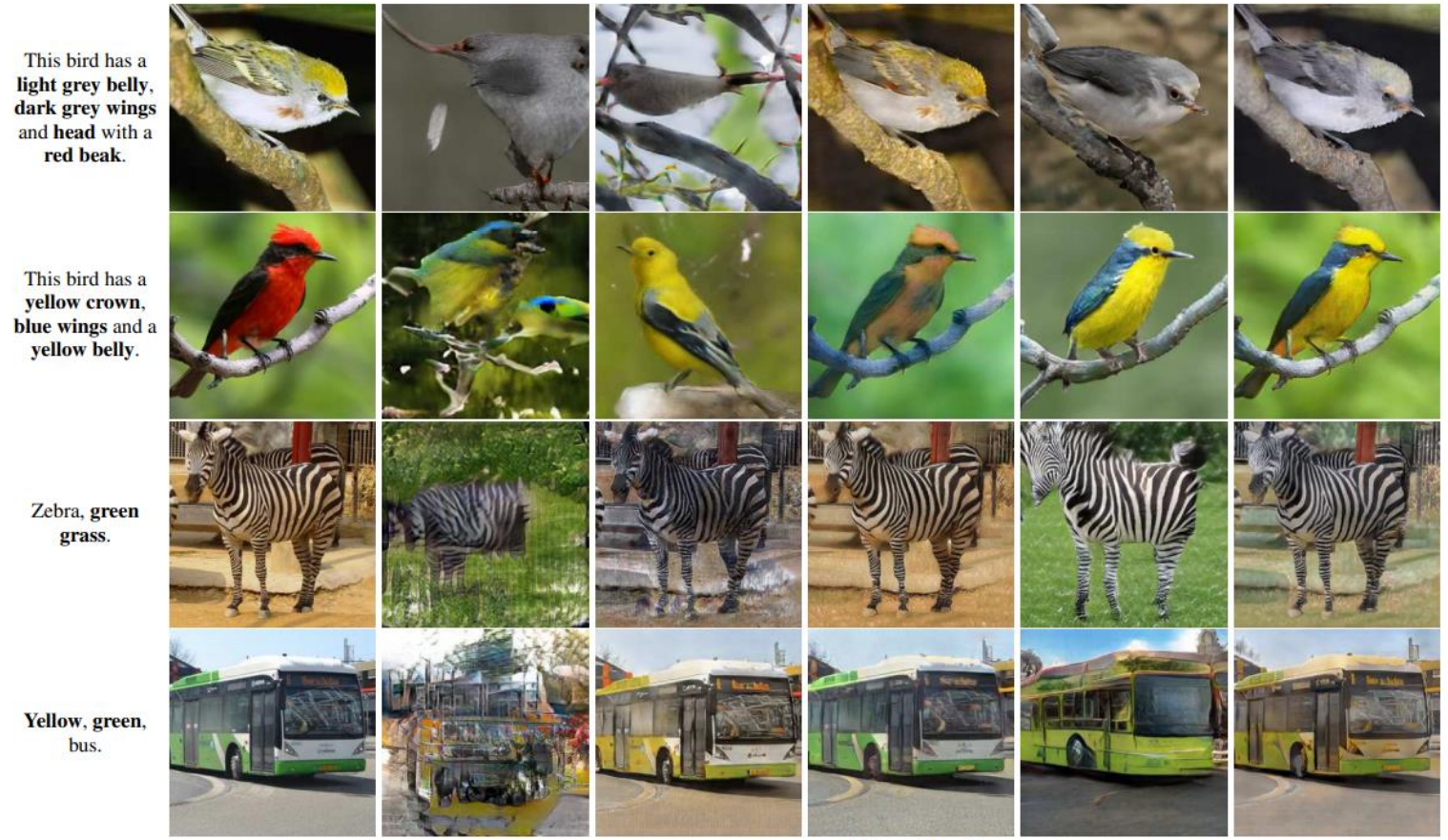


Figure 6: Qualitative comparison of three methods on the CUB bird and COCO datasets.

Experiment


This bird has a **light grey belly**, **dark grey wings** and **head** with a **red beak**.
This bird has a **yellow crown**, **blue wings** and a **yellow belly**.
Zebra, **green grass**.
Yellow, **green**, bus.
a: Text
b: Original
c: Ours w/o ACM
d: Our w/ Concat.
e: Ours w/o main module
f: Ours w/o DCM
g: Ours

Text-Guided Neural Image Inpainting

Lisai Zhang[1], Qingcai Chen*[1,2], Baotian Hu[1], Shuoran Jiang[1]
[1]Shenzhen Chinese Calligraphy Digital Simulation Engineering Laboratory, Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory, Shenzhen, China
LisaiZhang@foxmail.com, qingcai.chen@hit.edu.cn*, {baotianchina,shaunbysn}@gmail.com

MM '20: Proceedings of the 28th ACM International Conference on Multimedia

Motivation

PICNet

External guidance

Text guidance

Text guidance

Sentence-level，缺失了fine-grained

- Conditional GAN 模型，通过实验证明它具有根据文本描述合成合理性图片的性能。
- StackGAN将多个GAN网络stack起来，可以逐步生成高像素的图像。

word-level，含有fine-grained

- AttnGAN引出了对word-level 的注意力机制。该论文采用了AttnGAN中的matching loss。

The goal

1. **Fill the semantic information in corrupted images according to the provided descriptive text.**
2. **Keeping coherence with surroundings**

Origin | Corrupted | Without Text | **Text-guided (a)** | **Text-guided (b)** | **Text-guided (c)**

Text (a): *"This bird has gray head, black and white wings and yellow belly."* **Text (b)**: *"This bird is gray in color, with red belly."*
Text (c): *"A small bird with spotted blue wings, gray tail and head, and has white throat, breast, belly and white undertail."*

Application:

- Restoration of damaged paintings
- Photo editing
- Image rendering

Prior method

Prior methods	Upside	Downside
Diffusion-based Patch-based	Produce high-quality images	Fail in complicated scenes such as unique masks on objects or large holes
Deep learning-based	Can be in complicated scenes such as unique masks on objects or large holes	Cannot achieve high image appearance quality when filling irregular holes.

Challenge

1. image and text are heterogeneous; it is hard to transform image and text features to a shared space.

2. the descriptive text usually contains redundant information.

3. the model must distinguish between the information about the corrupted region and the remaining parts.

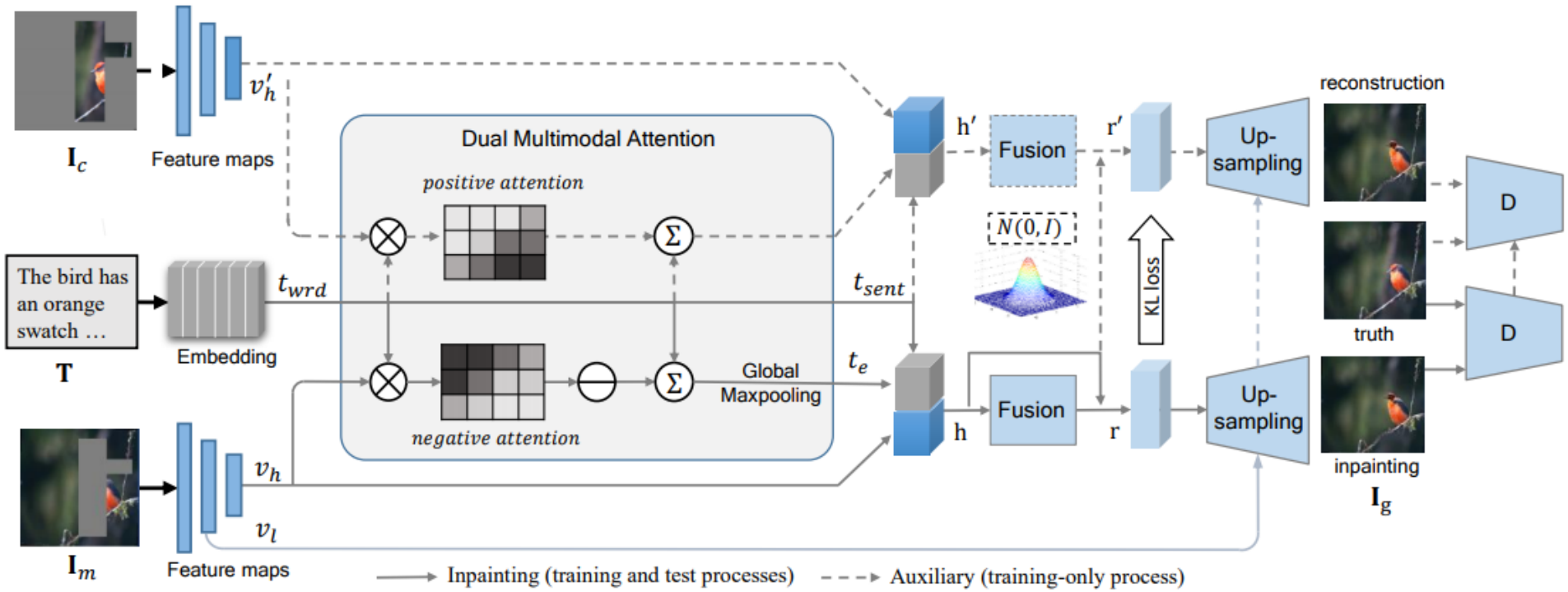
Model:
Text-Guided Dual Attention Inpainting Network (TDANet)

Step:

1. A dual multimodal attention mechanism is designed to extract the explicit semantic information about the corrupted regions.

2. An image-text matching loss is applied to maximize the semantic similarity of the generated image and the text.

Architecture of the TDANet



How?

The DAMSM loss

$$S(I,T) = \log\left(\sum_{i=1}^{L-1} \exp(\gamma cos(I_i, T_i))\right)^{\frac{1}{\gamma}} \tag{1}$$

$$P(I,T) = \frac{\exp(\gamma_2 S(I_i, T_i))}{\sum_{j=1}^{B} \exp(\gamma_2 S(I_i, T_j))} \tag{2}$$

$$DAMSM_w = -\sum_{i=1}^{B} (\log P(I,T) + \log P(T,I)) \tag{3}$$

Encoders for Image and Text

Image encoder

7-layers ResNet,

Use feature of the top layer to inference

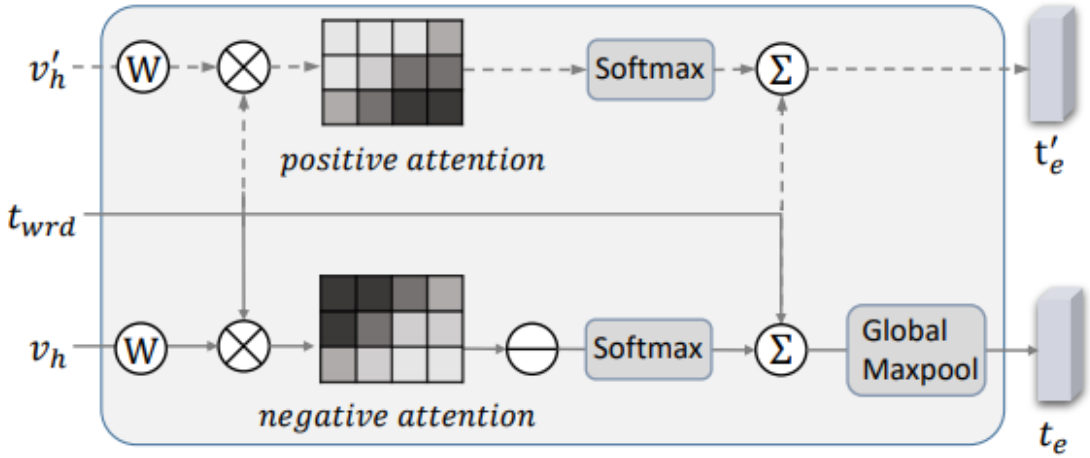
Use feature of the second last layer to reconstruct

Text encoder

GRU

t_{wrd}、t_{sent}

Dual Multimodal Attention Mechanism



$$s'_{i,j} = M'_i Q(v'_{hi})^T t_{wrdj} \tag{4}$$

$$s_{i,j} = -Q(v_{hi})^T t_{wrdj} + M_i \tag{5}$$

$$\beta_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N} \exp(s_{i,j})} \tag{6}$$

$$t_{ei} = \sum_{j=1}^{L} \beta_{i,j} t_{wrdj} \tag{7}$$

Experiments

datasets

Dataset	CUB	COCO
Mask Area (Avg)	34.59%	19.62%
Vocabulary size	5,450	27,297
Caption length (Avg)	15.23	10.45
Objects per image (Avg)	1	7.3
Captions per image	10	5

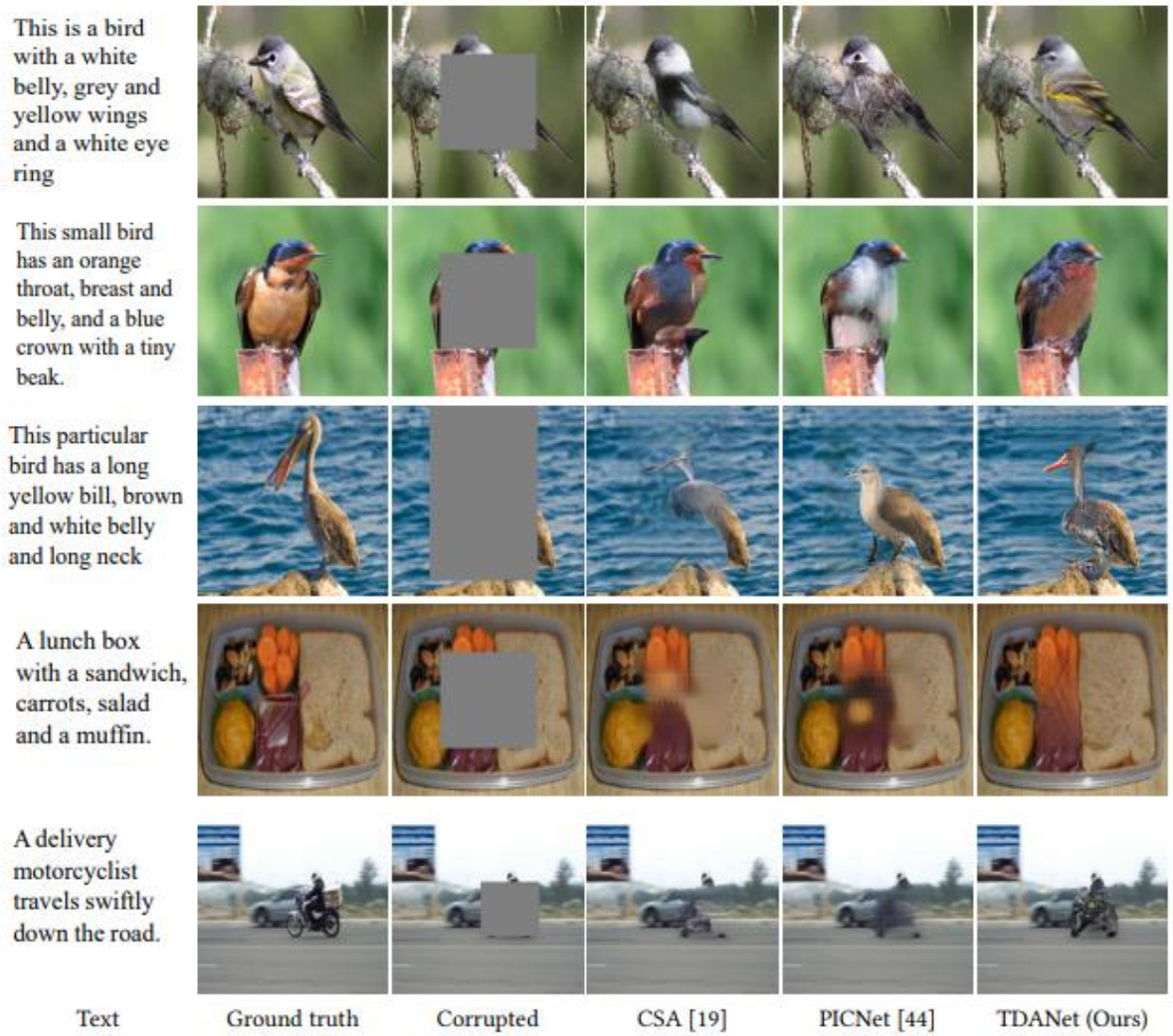
Experiments

Quantitative results

Dataset	Model	$\ell_1^-(\%)$			PSNR$^+$			TV loss$^-$(%)			SSIM$^+$(%)		
		center	object	$\|\|\Delta\|\|$	center	object	$\|\|\Delta\|\|$	center	object	$\|\|\Delta\|\|$	center	object	$\|\|\Delta\|\|$
CUB	CSA [19]	3.99	6.42	2.43	20.79	19.13	1.66	3.64	4.33	0.69	82.69	72.65	10.4
	PICNet [44]	3.65	7.28	3.63	20.96	18.80	2.16	3.71	4.12	0.41	84.51	70.78	13.73
	TDANet	**3.53**	**4.80**	**1.27**	**21.30**	**20.89**	**0.41**	**3.55**	**3.34**	**0.21**	**84.63**	**79.16**	**5.47**
COCO	CSA [19]	5.07	8.78	3.71	20.07	19.23	0.84	4.19	4.86	0.67	83.21	75.22	7.99
	PICNet [44]	5.53	9.21	3.68	19.57	18.73	1.02	4.51	4.97	0.46	81.78	74.44	7.34
	TDANet	**4.08**	**7.48**	**3.40**	**21.31**	**20.57**	**0.74**	**4.54**	**4.20**	**0.34**	**83.87**	**76.78**	**7.09**

Experiments

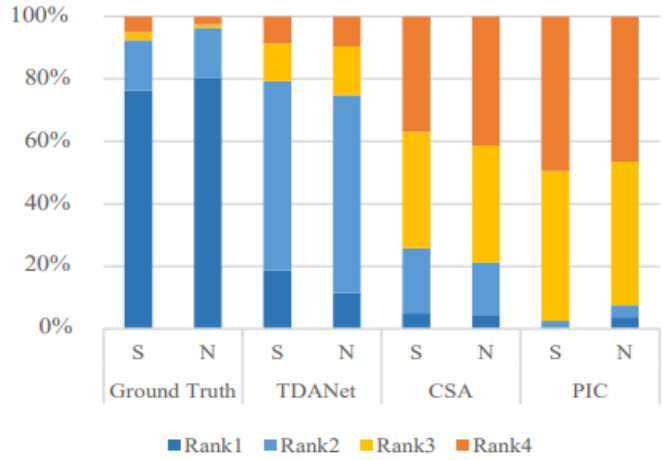
Qualitative results



Experiments

User study

Model	Naturalness	Semantic Consistency
Ground Truth	1.208	1.363
CSA	2.981	3.060
PICNet	3.178	3.469
TDANet	**2.531**	**2.106**



Conclusion

- 实验证明:
- A dual multimodal attention mechanism and an image-text matching based loss对该项任务是有很大改进的。

- 思考:

1. 在编码阶段，使用的模型比较简单，将来是否可以使用更复杂一点的模型。
2. 实验在COCO数据集上并没有取得很好的效果，可以使用一些数据增强技术来提高效果。
3. 是否可以使用模型生成过后的数据来和对词向量就行插值来扩充数据集，从而达到迭代的效果。

Efficient Neural Architecture for Text-to-Image Synthesis

Douglas M. Souza, Jônatas Wehrmann, Duncan D. Ruiz
School of Technology
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brazil

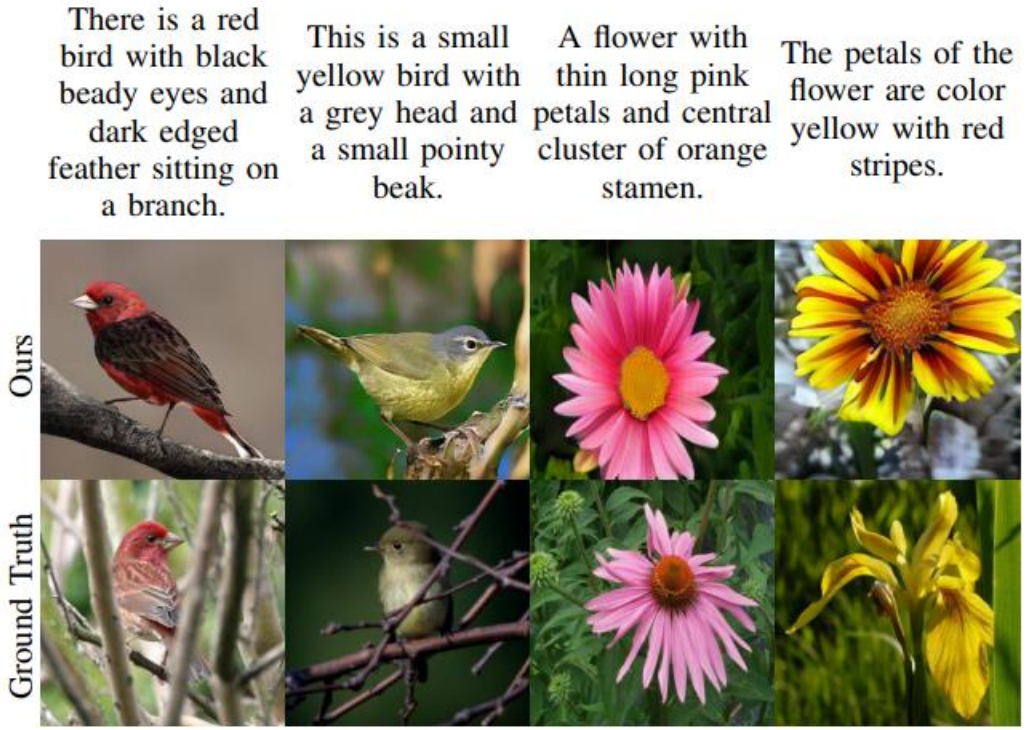
2020 International Joint Conference on Neural Networks (IJCNN)

Motivation

- It's difficult to combine data from two different modalities
- Most of methods have adopted a multi-stage training strategy

The goal

A task of generating images from text descriptions.



How?

Steps

1. encoding text descriptions to a vector representation

2. using this representation as a condition to train a Conditional GAN

Text Encoder

- encoding text descriptions to a vector representation by a pretrained DAMSM modual.

- Original image captions S are tokenized, vectors feed a Bidirectional GRU network，output per-token hidden representation、a global vector

Sentence Interpolation

Method:

use all the available captions for computing the general sentence embedding.

Advantage：

(i) it makes the sentence embedding space to be more smooth;

(ii) and also works as a data augmentation strategy.

Comparison

- Dataset：

Synthesis	Inpainting	Manipulation
CUB	CUB	CUB
Oxford-102	COCO	COCO

Comparison

- Challenges：

Synthesis	Inpainting	Manipulation
Cross modalities	**Fail in complicated scenes，like large hole、unique masks on objects**	**Cannot precisely correlate fine-grained words with corresponding visual attributes**
Multi-stage training causes complexity	**Cannot achieve high image appearance quality when filling irregular holes**	**Cannot effectively identify text-irrelevant contents**

Comparison

- Evaluation：

Synthesis	Inpainting	Manipulation
IS	**L1 loss**	**IS**
FID	**PSNR**	**Sim**
-	**TV**	**Diff**
-	**SSIM**	**MP**

Comparison

- Methods：

Synthesis	Inpainting	Manipulation
a single stage	**Encoders for image and text**	ACM
sentence interpolation	**Dual multimodal attention mechanism**	DCM

Conclusion

1. 针对该多模态任务，目前我感觉改进的点在于:
2. 卷积操作的改进：dilated convolution、gated convolution、partial convolution等一系列改进
3. 对文本信息进行加工处理，获取更fine-grained 的word信息、去除与任务不相关的文本信息。
4. 将文本信息和图片信息连接起来进行改进，使用各种注意力机制，如dual multimodal attention、ACM等

Thanks