Supplmentary Material: L-CoDer: Language-based Colorization with Color-object Decoupling Transformer

Zheng Chang[#1], Shuchen Weng[#2], Yu Li[3], Si Li[*1], and Boxin Shi[2]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2] NERCVT, School of Computer Science, Peking University
[3] International Digital Economy Academy
{zhengchang98,lisi}@bupt.edu.cn
{shuchenweng,shiboxin}@pku.edu.cn
liyu@idea.edu.cn

6 Appendix

6.1 Parameter Setting

We build three variants of our model to analyze the impact of different settings for the color-object decoupling transformer. We present the parameter settings and qualitative results of corresponding variants in Tab. 3 and Tab. 4. As can be seen, the performance improves with the increasing number of parameters. In the main paper, we report the results of L-CoDer (Large).

Table 3. Different parameter settings for model variants.

Model	Blocks L	Hiden size C_z	MLP size	Heads	Params
L-CoDer (Small)	4	768	3072	12	28M
L-CoDer (Base)	8	768	3072	12	57M
L-CoDer (Large)	12	1024	4096	16	151M

Table 4. Qualitative results of model variants. ↑ (↓) means higher (lower) is better.

Model	SSIM↑	LPIPS↓	LPIPS↓
L-CoDer (Small)	25.003	90.942%	0.171
L-CoDer (Base)	25.325	91.585%	0.161
L-CoDer (Large)	**25.504**	**91.963**%	**0.159**

[#] Equal contributions. * Corresponding author.

6.2 Visualization of Applying OCCM

We visualize the attention maps before/after applying OCCM (Fig. 9 left) and OCCM (Fig. 9 right) at layers 3, 6, and 9. As illustrated in Sec. 3.4 of the main paper, the attention maps before applying OCCM are calculated by noun tokens and image tokens to locate objects, which are transferred by the OCCM to guide adjective tokens to colorize image tokens.

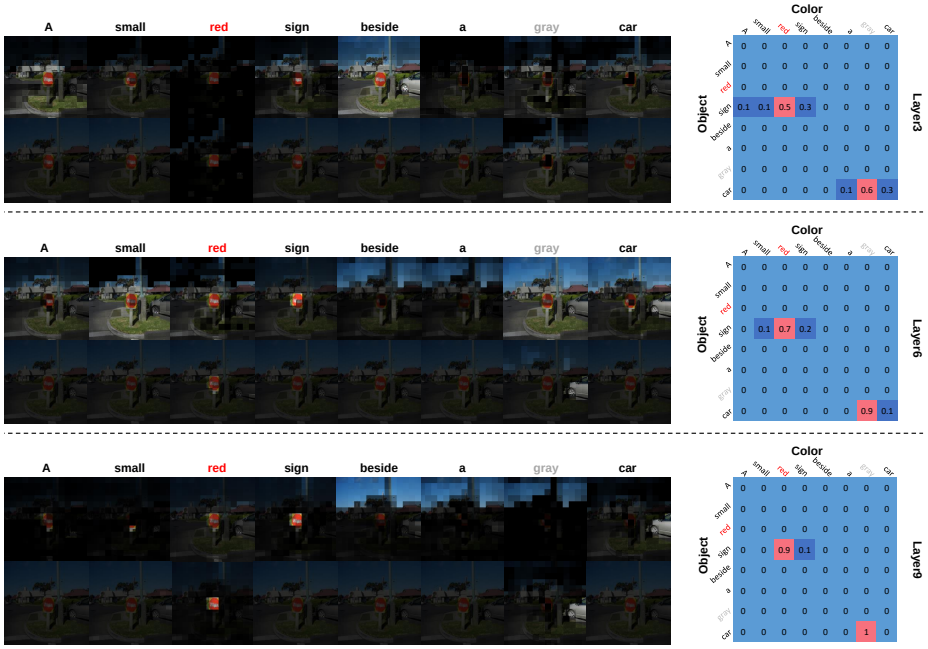


Fig. 9. **Left top/bottom**: Attention maps before/after applying the OCCM [8]. The brighter the region in the visualization, the greater the probability that the object is located in that region. As the number of layers increases, attention maps before applying the OCCM gradually find object positions corresponding to noun tokens (*e.g.*, "sign" and "car" tokens), where mismatched positions (*e.g.*, corresponding to "beside" and "a") are filtered by the OCCM. As a result, attention maps after applying OCCM reveal the corresponding relationship between the image tokens and adjective tokens, which ensures colors could be correctly applied to the corresponding regions. **Right**: OCCM. Numbers in the matrix show the probability that noun tokens and adjective tokens are in the same combination. As the number of layers increases, the OCCM predicts combinations more accurately.

6.3 Additional Quantitative and Qualitative Results

We qualitatively compare our L-CoDer with four automatic methods, *e.g.*, CIC [12], Deoldify [1], InstColor [5], and ChromaGAN [7]. We also make comparisons with four language-based methods, *e.g.*, LBIE [2], ML2018 [4], Xie2018 [9], and L-CoDe [8]. We further create four baselines to demonstrate the effectiveness of decouple, bidirection, evolution, and upsample, whose details are described in Sec. 4.3 of the main paper.

To draw a more comprehensive evaluation of L-CoDer and other comparison methods mentioned above, we additionally use two quantitative metrics to measure the synthetic image quality, *i.e.*, R-precision (R-prcn) [10] and Fréchet inception distance (FID) [3]. R-precision is used to evaluate whether colorized images are well conditioned on the given language condition. FID [3] is a perceptual similarity metric to evaluate the distance between the generated images and original images with VGG backbone. Quantitative comparison and ablation experiments with additional metrics are shown in Tab. 5, where our method performs best on both metrics.

We show more experimental comparison results in Fig. 10 and Fig. 11. We provide additional demonstrations of L-Coder's advantages, *i.e.*, unified modalities, accurate color representation, and local robustness, in Fig. 12 (corresponding to Fig. 1 of the main paper). We also show more ablation study results in Fig. 13.

Table 5. Quantitative comparison result. L-CoDer (ours) performs best in two metrics. ↑ (↓) means higher (lower) is better.

Category	Method	R-prcn ↑	FID ↓
Automatic	CIC [12]	41.758%	30.841
	DeOldify [1]	42.598%	30.471
	InstColor [5]	42.397%	30.500
	ChromaGAN [7]	42.920%	33.834
Language-based	LBIE [2]	42.276%	32.594
	ML2018 [4]	43.443%	33.908
	Xie2018 [9]	41.954%	33.137
	L-CoDe [8]	44.046%	30.718
Ablation	W/o decouple	42.437%	33.685
	W/o evolution	44.086%	33.323
	W/o bidirection	43.282%	33.179
	W/o upsample	42.316%	30.425
Ours	L-CoDer (Small)	43.161%	31.829
	L-CoDer (Base)	43.322%	30.600
	L-CoDer (Large)	**47.103%**	**30.097**

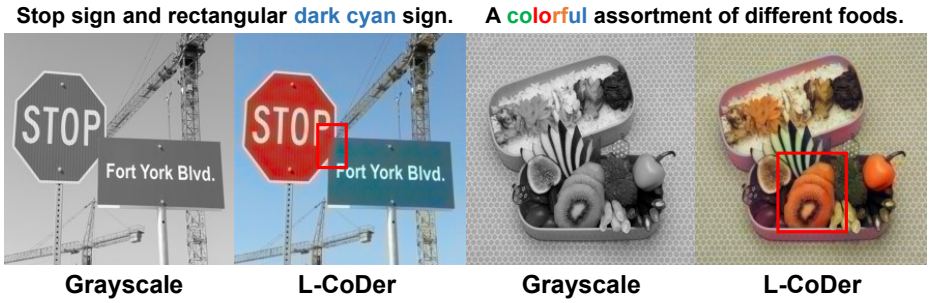
6.4 Additional Ablation

We conduct "W/o decouple" ablation experiment by removing all the decoupling modules and BCE loss to optimize the predicted OCCM. To further explore whether "the loss alone is leading to the improved performance", we design "w/o BCE" ablation variant by removing BCE loss while remaining the decoupling modules. As shown below, both the decoupling modules and BCE loss improve the evaluation scores and alleviate the color-object mismatching problem (the yellow on the big bin fades). It should be noted that the model must be equipped with the decoupling modules to adopt the BCE loss. Otherwise, there are no noun tokens Z_{obj} and adjective tokens Z_{col} to predict OCCM M_{occm} as the optimized term. Therefore, it is impossible to design an ablation variant without decoupling modules that can predict OCCM to calculate BCE losses.

Method	PSNR↑	SSIM↑	LPIPS↓	R-prcn ↑	FID ↓
W/o decouple	25.014	90.724%	0.173	42.437%	33.685
W/o BCE	25.179	91.029%	0.169	44.569%	32.688
Ours	**25.504**	**91.963%**	**0.159**	**47.103%**	**30.097**

6.5 Failure Cases

We show two failure cases below. In the left case, there are color bleeding between the two signs which is most likely caused by the adherent boundary at the small overlapping region. On right, the kiwi fruit is colorized as orange under the guidance of the adjective "colorful". This is probably because our model could not correctly recognize the semantics of some rare and small objects.



6.6 Additional user study

We conduct an additional user study to enrich our experiments using the following setting: We mix generated images from models and the real image, and ask participants to choose the one that they think is real. This experiment also follows the protocol in Sec. 4.2. As the results shown below, for the question "what is real", the pick rate of our model is very close to the ground truth, which is the highest among related methods.

LBIE [2]	ML2018 [4]	Xie2018 [9]	L-CoDe [8]	Ours	Ground truth
9.64%	11.44%	14.20%	15.60%	22.24%	26.88%

6.7 Model Size

We list the model size of other compared methods and L-CoDer variants below. Note that the number of upstream task parameters is not counted, *e.g.*, , detection model in InstColor [5] and pre-trained encoders in Xie2018 [9], L-CoDe [8] and L-CoDer variants.

Automatic	CIC [12]	DeOldify [1]	InstColor [5]	ChromaGAN [7]
	32M	218M	34M	174M
Language-based	LBIE [2]	ML2018 [4]	Xie2018 [9]	L-CoDe [8]
	14M	18M	50M	21M
Ablation	W/o decouple	W/o evolution	W/o bidirection	W/o upsample
	132M	134M	151M	147M
Ours	L-CoDer (Small)	L-CoDer (Base)	L-CoDer (Large)	
	28M	57M	151M	

6.8 Application

We demonstrate the controllability of L-CoDer by colorizing photos with different captions in Fig. 14. We show more results of colorizing legacy photos in Fig. 15 to demonstrate our generalization capability.

6.9 Upsampler Structure Details

We describe details of upsampling layers (corresponding to Sec. 3.6 of the paper) in Tab. 6, where Deconv means deconvolution [11] and IN means instance normalization [6].

Table 6. Detailed architecture of the upsampling layers.

Block	Operation	Channel number	Kernel size	Stride	Padding	Activation	Norm
Up-1	Deconv	768	4	2	1	ReLU	-
Conv-1	Conv	256	3	1	1	ReLU	IN
Up-2	Deconv	256	4	2	1	ReLU	-
Conv-2	Conv	128	3	1	1	ReLU	IN
Up-3	Deconv	128	4	2	1	ReLU	-
Conv-3	Conv	64	3	1	1	ReLU	IN
Up-4	Deconv	64	4	2	1	ReLU	-
Output	Conv	32	3	1	1	tanh	-

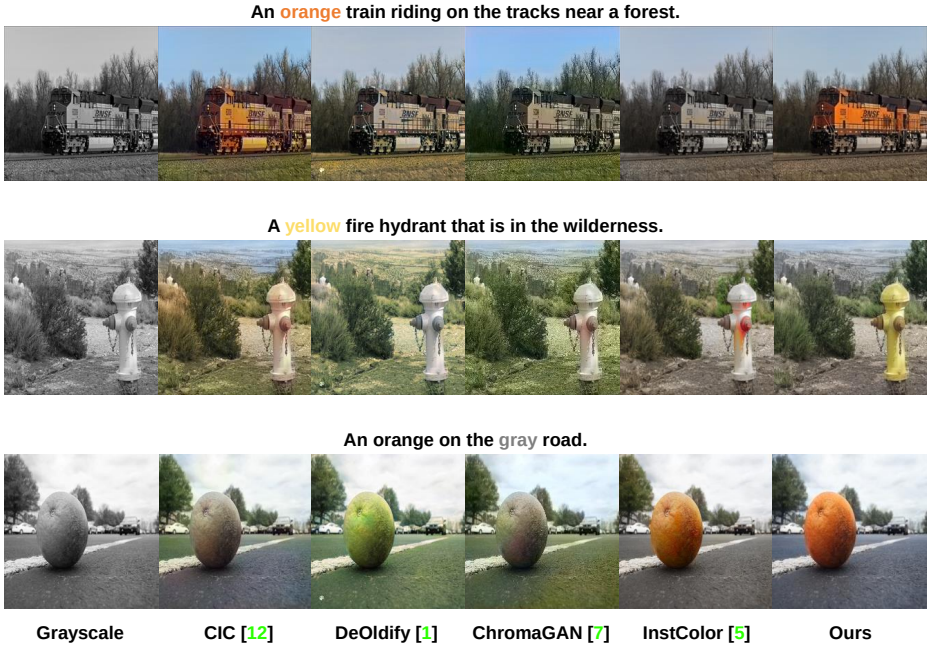


Fig. 10. More comparison results with CNN-based methods.

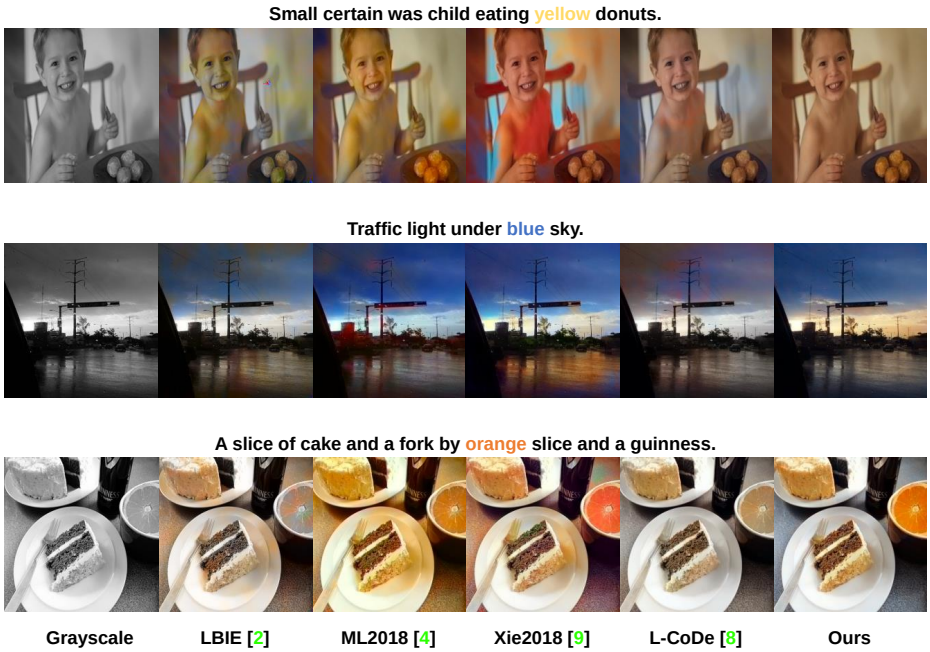


Fig. 11. More comparison results with Transformer-based methods.

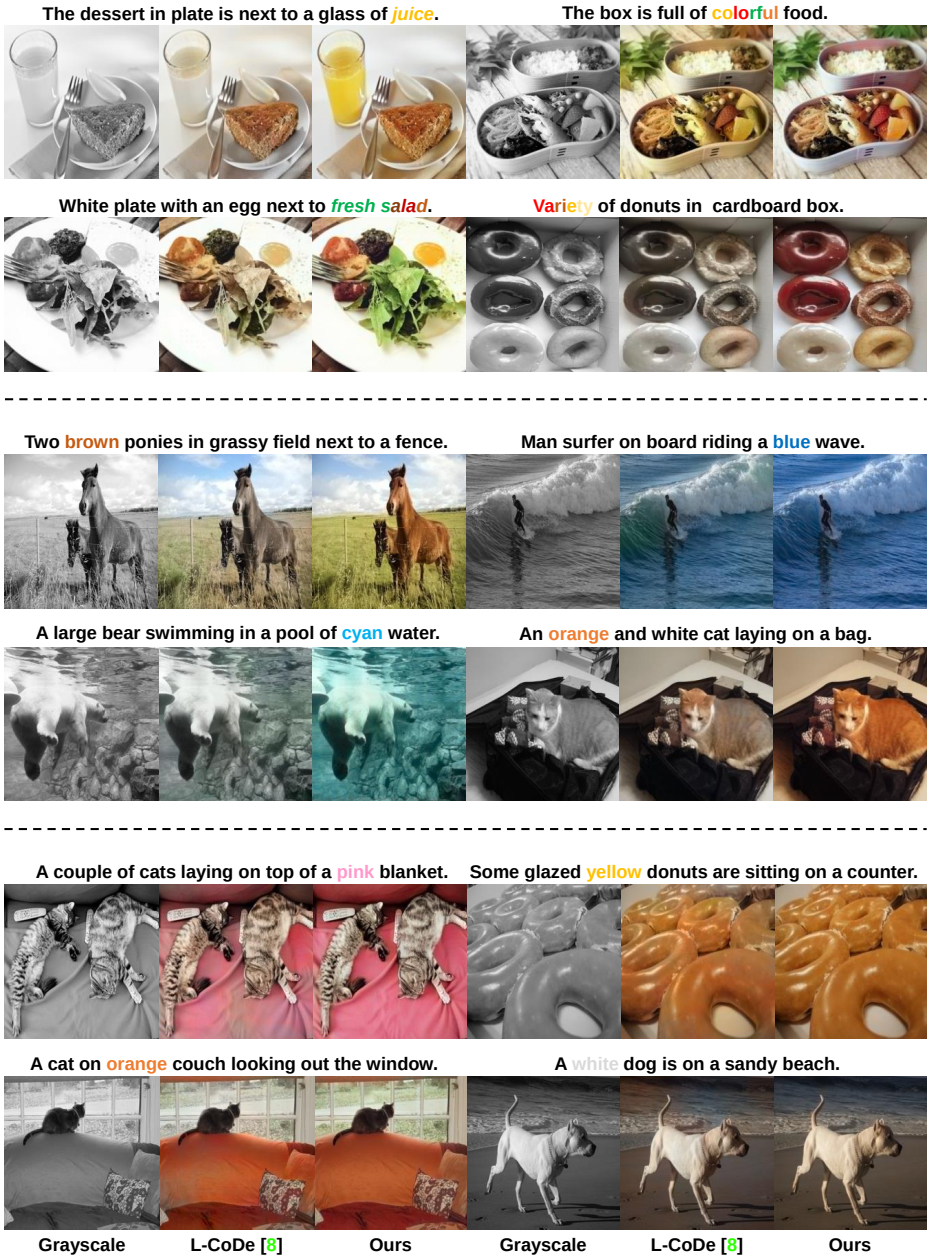


Fig. 12. Additional demonstrations of L-Coder's advantages. **Row 1&2**: our method understands intrinsic color properties behind the word. **Row 3&4**: our method generates accurate and plausible colors. **Row 5&6**: our method has stronger robustness to locally strong variation of texture or luminance.

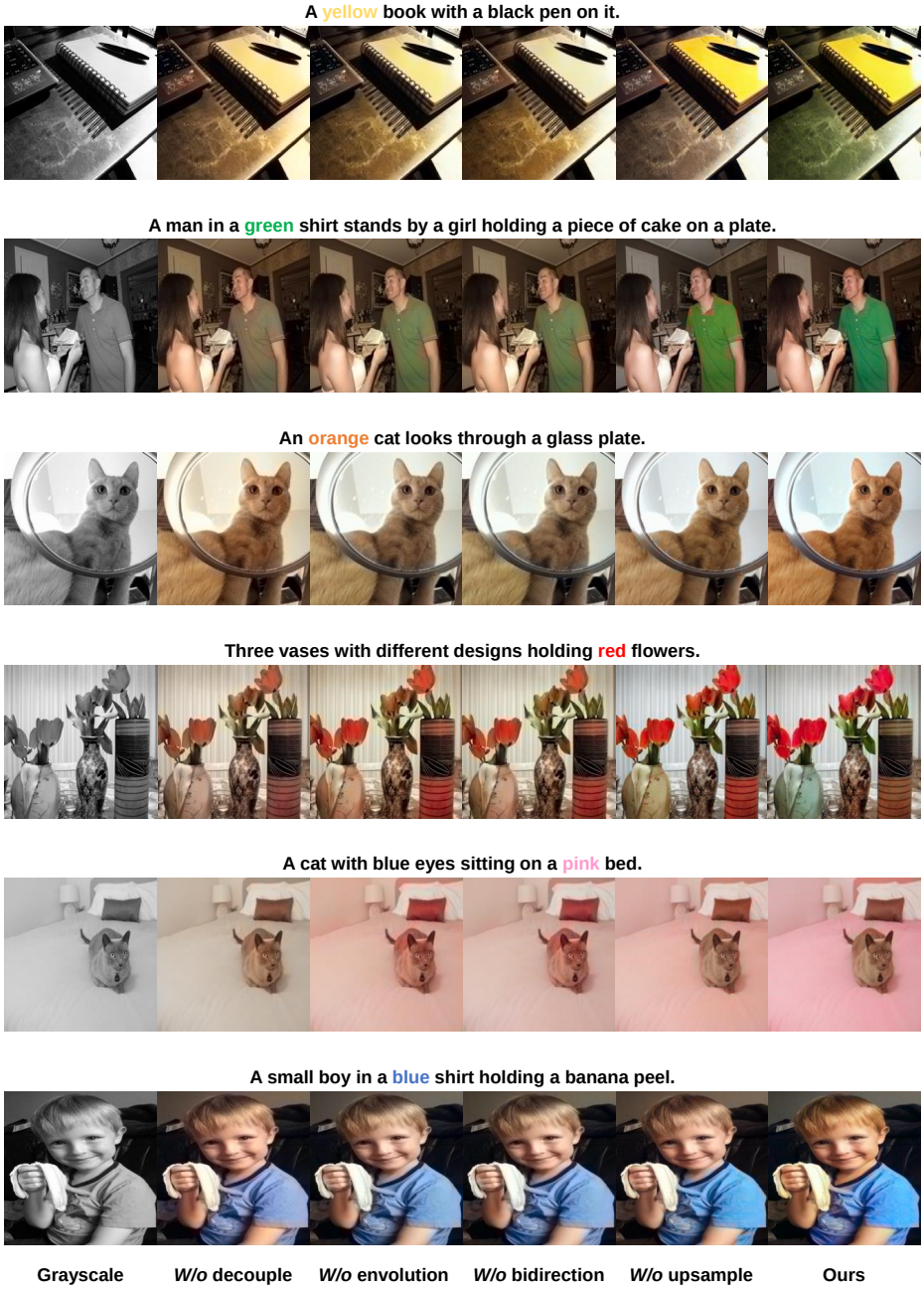


Fig. 13. More ablation study results.

Fig. 14. Diverse colorization results with different captions.

Fig. 15. More colorization results of legacy photos.

References

1. Antic, J.: A deep learning based project for colorizing and restoring old images (and video!), https://github.com/jantic/DeOldify 3, 5
2. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-based image editing with recurrent attentive models. In: CVPR (2018) 3, 5
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. In: NIPS (2017) 3
4. Manjunatha, V., Iyyer, M., Boyd-Graber, J., Davis, L.: Learning to color from language. In: NAACL (2018) 3, 5
5. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: CVPR (2020) 3, 5
6. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 6
7. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: WACV (2020) 3, 5
8. Weng, S., Wu, H., Chang, Z.C., Tang, J., Li, S., Shi, B.: L-code: Language-based colorization using color-object decoupled conditions. In: AAAI (2022) 2, 3, 5
9. Xie, Y.: Language-guided image colorization. Master's thesis, ETH Zurich, Departement of Computer Science (2018) 3, 5
10. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018) 3
11. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014) 6
12. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016) 3, 5