RORD: A Real-world Object Removal Dataset

Min-Cheol Sagong*
mcsagong@dali.korea.ac.kr
Yoon-Jae Yeo*
yjyeo@dali.korea.ac.kr
Seung-Won Jung
swjung83@korea.ac.kr
Sung-Jea Ko
sjko@korea.ac.kr

Korea University
Seoul, Korea

Abstract

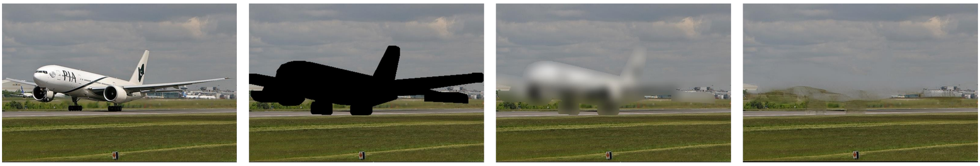
Various convolutional neural networks (CNNs)-based image inpainting techniques have been actively studied to remove unwanted objects or restore missing parts in recent years. The common standard for training image inpainting CNNs is synthesising hole regions on the existing datasets, such as ImageNet and Places2. However, from the viewpoint of the object removal task, such a methodology is suboptimal because actual pixels behind objects, *i.e.*, "ground truth", cannot be used for training. Facing this problem, we introduce Real-world Object Removal Dataset (RORD), a large-scale collection of image pairs with and without objects. RORD consists of a wide range of real-world scenes, plus two types of pixel-accurate annotations, *i.e.*, object mask and segmentation map. Our dataset allows existing image inpainting models to be trained accurately as well as evaluated with high confidence. In this paper, we describe in detail how the dataset is constructed and demonstrate the validity and usability of RORD. RORD is publicly available at `https://github.com/Forty-lock/RORD`

1 Introduction

In recent years, research attention towards image inpainting for image restoration and object removal has grown faster. Unlike classic image inpainting methods that require human labour or expert knowledge to complete missing image content, recent convolutional neural network (CNN)-based methods have made it possible to reconstruct plausible pixels automatically. Advanced image inpainting technology has already come close to our lives.

Existing deep learning-based image inpainting methods [7, 13, 14, 15, 20, 22, 26, 27, 28] require pairs of inpainting masks and ground truth images with the pixel values in the masks for training. A common practice is creating a hole region of the desired size and shape on the image and using the original image as the ground truth for supervision. Existing image datasets, such as ImageNet [8] and Places2 [29], have thus been used to generate a training

*Equal Contribution.



(a) (b) (c) (d)

Figure 1: Toy example for the object removal task: (a) Input image containing the object to be removed (plane), (b) object-removed image, (c) and (d) object removal results. The PSNRs are obtained as 21.10 and 18.69 for (c) and (d), respectively. This inconsistency misleads the training and performance evaluation of inpainting models.

dataset. However, this practice has an obvious drawback when applied to the object removal task. Object removal aims to erase the specified object and fill the hole region such that the region blends seamlessly with the surrounding background. Under the common practice, the ground truth pixels for the hole region are still object pixels. Consequently, inpainting networks are unavoidably trained to synthesise these object pixels, which is not desired for the target task of object removal. Figure 1 depicts a toy example of object removal. Figures 1(c) and (d) show two different inpainting results for the input image and object mask in Figures 1(a) and (b). It is evident that Figure 1(d) is better than Figure 1(c) from the perspective of object removal. However, in terms of the standard performance measure of the peak signal-to-noise ratio (PSNR), Figures 1(c) and (d) have 21.10 dB and 18.69 dB, respectively. In other words, Figure 1(c) is evaluated to be a better result than Figure 1(d), misleading not only the model training but also performance benchmarking.

This paper introduces the Real-world Object Removal Dataset (RORD), which is the first large-scale real-world image dataset specialised for the object removal task. To the best of our knowledge, RORD is the only dataset that contains a sufficient number of images with and without the objects in the scene. RORD consists of 516,709 images captured under 3,447 unique scenes. Each scene belongs to one of 55 outdoor and 32 indoor categories and has the corresponding ground truth image without objects to be removed. Moreover, two types of pixel-wise annotations are provided for all images: binary object masks and semantic segmentation maps consisting of 42 classes. The images and abundant annotations of RORD are publicly available for researchers to accelerate their studies.

Our contributions are summarised as follows:

- We release RORD, the first real-world image dataset that is specialised for object removal, containing a pair of images with and without objects.
- RORD supports a large number of images captured under a wide variety of indoor and outdoor real-world environments and contains objects of various sizes and classes.
- The two types of pixel-accurate annotations, *i.e.*, object masks and segmentation maps, are further provided to advance the field of image inpainting and other related tasks.

2 Preliminaries

2.1 Image Inpainting

Image inpainting techniques can be divided into two approaches: traditional and learning-based approaches. In the traditional approach, diffusion-based methods [2, 3, 10] propagate

Figure 2: Three examples of video clips in our RORD. For each clip, with corresponding segmentation maps, we present a ground truth image (first column) without objects as well as object-containing images.

the information from the neighbouring regions to the hole regions, while patch-based methods [1, 7, 9, 25] paste patches sampled from the background regions into the missing regions. Especially, Barnes *et al*. introduced a fast approximate nearest neighbour patch search algorithm, called PatchMatch [1]. However, the traditional methods have limited performance since they cannot consider semantic information or global structure.

On the other hand, the learning-based approach [13, 15, 17, 20, 22, 23, 26, 27, 28] aims to extract semantic information from massive data training, thus significantly enhancing the performance of image inpainting. Furthermore, the generative adversarial network (GAN) has been applied to recover the corrupted regions more plausibly, resulting in further improvements. Specifically, Pathak *et al*. presented ContextEncoder [17], an encoder-decoder model to to complete the hole region. Yu *et al*. [26] proposed a novel model with coarse-to-fine architecture that generates an initial coarse result and refines it based on the roughly filled prediction. Moreover, with the parallel extended-decoder path (PEPSI) [20], Sagong *et al*. improved the inpainting performance while reducing the number of convolution operations. Meanwhile, several adaptive convolution techniques considering pixels' validity were presented: partial convolution [13] and gated convolution [27]. Similarly, Yu *et al*. introduced a spatial region-wise normalisation called region normalisation (RN) [28], which normalises pixels in corrupted and uncorrupted regions separately and performs affine transformations globally.

Meanwhile, some studies aimed at erasing whole objects have been recently introduced [4, 8, 16, 19, 21]. Since there are no ground truth images with objects removed, they cannot consider reconstruction errors such as L_1 or L_2 loss. To determine the presence or absence of an object in the result image, Shetty *et al*. [21] employ an object classifier. SESAME [16] exploits semantic labels of the hole regions to synthesise the corresponding pixels. Similarly, SECI-GAN [19] not only extracts high-level cues from semantic segment informa-

Figure 3: Examples of the image pairs and segmentation maps with and without objects. In the first row, each image contains a person or airplane as a target object to be removed. The second row shows background images without target objects.

Charateristic		Number	Rate(%)
		516,705	100
Location	Outdoor	334,376	64.7
	Indoor	182,329	35.3
Object proportion	$\sim 10\%$	121,453	23.5
	$10\% \sim 20\%$	122,873	23.7
	$20\% \sim 30\%$	103,905	20.1
	$30\% \sim$	168,474	32.7

Table 1: Image distributions in terms of location and object proportion. RORD provides a sufficient number of images for each level, making it especially useful to train object removal models robust to scenes and object sizes.

tion, but also utilises the fine-grained details captured by edge extraction. Moreover, these methods could not evaluate their models with the conventional metrics such as PSNR and SSIM [24, 31].

2.2 Datasets for Image Inpainting

Various vision datasets can be employed for deep-learning-based image inpainting. Indeed, state-of-the-art image inpainting methods have used popular large-scale datasets developed for image classification, such as ImageNet [8] and Places2 [29]. By creating rectangular or free-form holes in the image, hole and mask images are obtained to be used as input for training or evaluation. The original image without holes serves as the ground truth. Although this procedure is simple as well as effective for general image inpainting, there is a glaring error in its application to remove dispensable objects. When the entire object belongs to the hole, the ground truth image still contains the object that cannot and should not be restored. These mismatched pairs lead to miscalculated losses resulting in inaccurate learning of models. Moreover, since there is no segmentation information, these datasets cannot be employed at all for the evaluation of object removal tasks. On the one hand, object segmentation datasets like MS-COCO [12] or cityscapes [6] easily derive hole images masking the object, but there are still no correct ground truth images with objects removed.

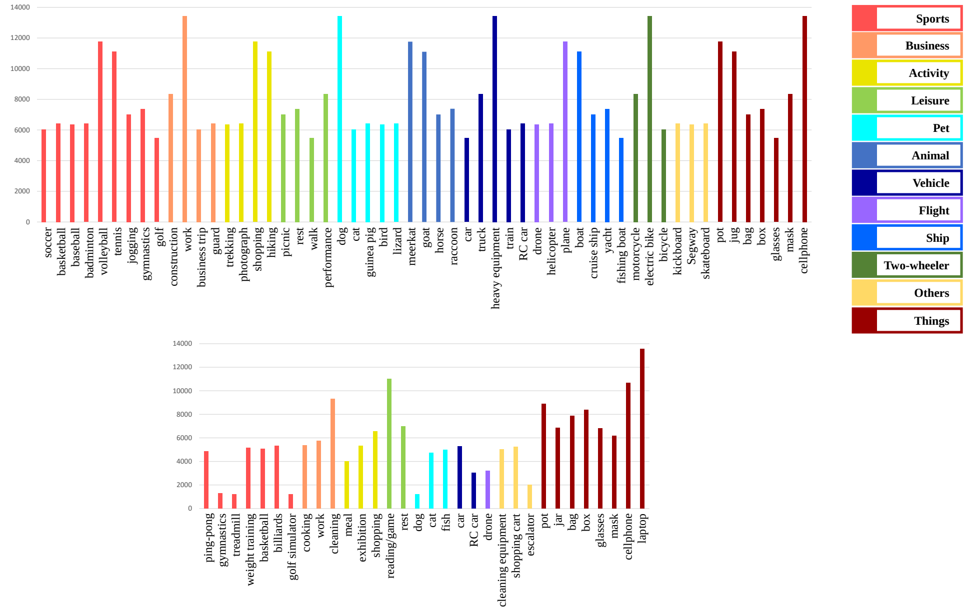


Figure 4: Image statistics in terms of scene categories for the outdoor (top) and indoor (bottom) environments. The Same colored bar means the same category. The outdoor dataset consists of 55 subcategories for 12 scene categories. Several infeasible categories are excluded for the indoor environments, resulting in 32 subcategories for 9 scene categories.

3 Real-world Object Removal Dataset

3.1 Dataset Statistics

RORD consists of 516,705 images captured under 3,447 unique scenes. Each scene was first captured without any target objects, serving as a ground truth image for object removal. Then, the same scene was captured multiple times with the target objects at different positions. In this manner, RORD provides real image pairs with and without objects, which are vital for the supervised learning of object removal models. We collected full HD videos and equalised their resolution to 1920×1080. The high-resolution images of RORD leave a greater room for posterior data augmentation. Figure 2 shows three scenes in RORD. As can be seen, we captured controlled scenes without any camera motion and located target objects at multiple positions such that multiple image pairs can be provided for each scene. Figure 3 shows more examples of the images with and without the objects. To maximise the diversity of the dataset, we define 12 scene categories: Sports, business, activity, leisure, pet, animal, vehicle, flight, ship, two-wheeler, things, and others. Figure 4 shows image distributions for outdoor and indoor environments. The outdoor dataset consists of 55 subcategories for 12 scene categories. Several infeasible categories are excluded for the indoor environments, resulting in 32 subcategories for 9 scene categories. In total, RORD provides 334,376 outdoor and 182,329 indoor images, which are sufficient to train object removal networks.

In addition, RORD supports images composed of objects of various sizes. The object size is classified into four different levels: ~10%, 10~20%, 20~30%, and 30%~, according

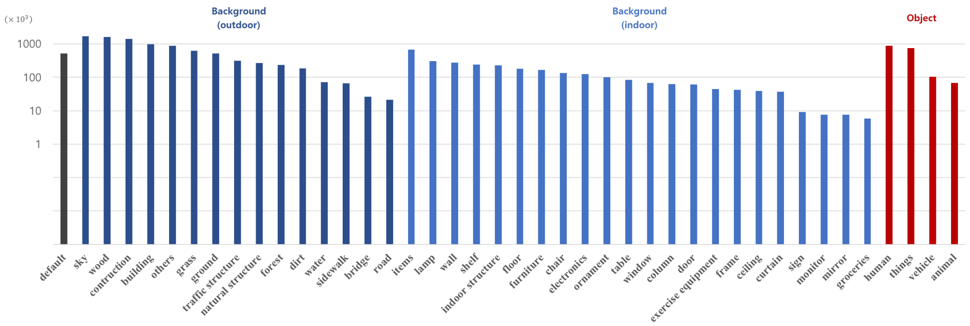


Figure 5: Number of pixels per annotation labels, which are grouped by scene categories and sorted according to the frequency for each category. The pixel distribution of semantic classes is widely spread over diverse classes.

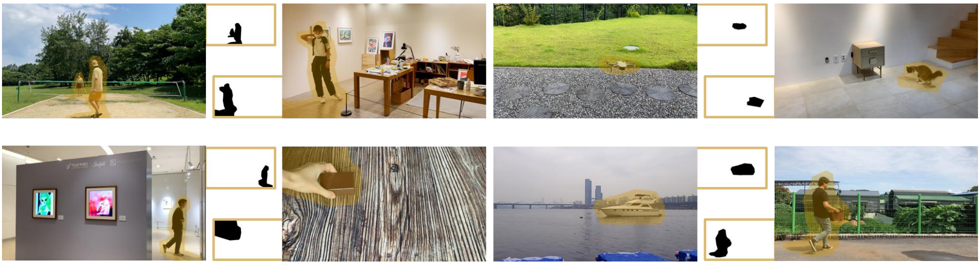


Figure 6: Examples of the object masks in RORD. We annotate the object with an enough margin for each image (as highlighted in yellow) and generate a binary mask to cover the object and artifacts from it completely.

to the proportion of the number of pixels in the object over the number of pixels in the image. As shown in Table 1, RORD provides a sufficient number of images for each level, making it especially useful to train object removal models robust to object sizes.

Last, to boost the performance of deep neural networks, elaborate data augmentation or post-processing is frequently applied. To support any desired data processing, we distribute full HD images without applying any post-processing. We intend to leave the choice of optimal handling of our images to researchers who deploy our dataset.

3.2 Annotations

RORD provides binary object masks to indicate object pixels to be removed, which are usually assumed as given in the image inpainting task. In general, object mask is generated from the semantic segmentation label. However, the object mask from the segmentation map does not completely cover the object and artifacts from it. For complete object removal, not only the objects but also their reflection and shadow need to be annotated. As shown in Figure 6, our object masks cover whole objects with proper margins. By using the object mask in RORD, object removal models can be evaluated properly.

Having pixel-wise semantic labels extends the feasibility of the dataset by allowing the development of various vision applications and multi-modal tasks. In this regard, RORD

		MS-COCO		Places2		RORD	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Deepfill-v1 [26]	PASCAL-VOC	17.01	0.774	17.10	0.777	16.87	0.773
	MS-COCO	19.17	0.832	19.25	0.835	19.07	0.831
	RORD	24.43	0.864	24.49	0.868	**24.81**	**0.869**
Partial Conv [13]	PASCAL-VOC	16.59	0.745	17.03	0.770	16.78	0.755
	MS-COCO	18.81	0.812	19.24	0.830	18.81	0.814
	RORD	24.00	0.854	24.39	0.860	**24.65**	**0.861**
Deepfill-v2 [27]	PASCAL-VOC	16.88	0.774	17.01	0.776	16.90	0.770
	MS-COCO	19.08	0.832	19.10	0.834	19.08	0.828
	RORD	24.46	0.865	24.43	**0.868**	**24.80**	0.866
PEPSI [20]	PASCAL-VOC	16.81	0.772	17.03	0.781	16.69	0.767
	MS-COCO	19.06	0.831	19.23	0.838	18.96	0.828
	RORD	24.45	0.865	24.80	0.838	**24.87**	**0.868**
RN [28]	PASCAL-VOC	16.31	0.759	16.44	0.744	16.54	0.757
	MS-COCO	18.53	0.823	18.59	0.813	18.68	0.820
	RORD	23.32	0.856	22.26	0.831	**23.91**	**0.858**

Table 2: PSNR and SSIM results of cross-validation test. Each model is respectively trained on MS-COCO, Places2 or RORD and evaluated on three other datasets, including PASCAL-VOC, MS-COCO, and RORD. The big numeric margin between the results evaluated with conventional methods or RORD is caused by the absence of ground truth data.

includes precise annotations, which can be divided into two types; 1) dynamic objects that appear in the scene or not, *e.g.*, humans, vehicles, and 2) static backgrounds that remain in every frame, *e.g.*, sky and ground. Accordingly, we separate annotation labels into two super labels, *i.e.*, objects and backgrounds, as depicted in Figure 5, covering 42 semantic classes selected from existing datasets, including MS-COCO [12], PASCAL-VOC [11], Places2 [29], and ADE20K [30]. Note that the pixel distribution of semantic classes is widely spread over diverse classes. Paid and experienced workers annotated all images, and unreliable annotations were manually excluded from the collections. We also reviewed the data by supervisors who are independent of the annotator to double-check for errors.

4 Evaluation

4.1 Evaluation Models

We tested the state-of-the-art image inpainting models, *i.e.*, Deepfill-v1 [26], Partial Conv [13], Deepfill-v2 [27], PEPSI [20], and RN [28], to evaluate the validity of RORD. Deepfill-v1 introduces the coarse-to-fine network and the contextual attention module to reconstruct the hole region using the patches in the background. Partial Conv applies a masked convolution with renormalisation to use only valid pixels, *i.e.*, the pixels outside the hole region. Deepfill-v2 utilises the gated convolution to better handle valid pixels for inpainting. PEPSI modifies the coarse-to-fine network into the parallel network to reduce the inference time. RN computes the mean and variance separately for valid and invalid pixels for normalisation to overcome the mean and variance shifts caused by the conventional feature normalisation.

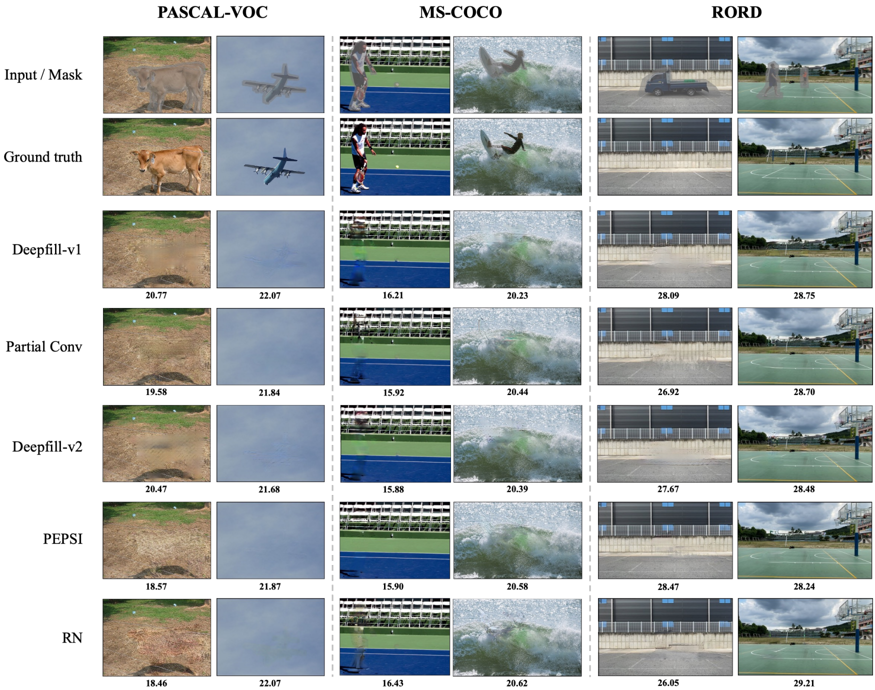


Figure 7: Visual comparison of results from various models trained with RORD. The numbers under the images represent PSNR between the result and the ground truth. RORD has higher reliability of performance measurement than the conventional datasets.

4.2 Implementation Details

We evaluated the aforementioned models trained on RORD, Places2 [29], or MS-COCO [12]. Specifically, since there is no background information behind the objects in Places2 and MS-COCO, we generated random object masks to train the models. On the contrary, for the RORD-trained models, we used pairs of images with and without objects. We divide RORD 412,304 images for training and 104,401 images for the test. For a fair comparison, we trained the inpainting models while keeping all settings unchanged.

The object-containing and object-less frames can have a slight misalignment and brightness shifts. For example, in an outdoor scene, brightness and background clutters can be changed by cloud or wind during the video clip. Therefore, we cropped the object region from the object-less image and pasted it to the object-containing image to alleviate the misalignment of the background. In addition, if the brightness of the images is different, simply pasting the object region can create unnatural boundaries. To cope with this problem, we applied the Poisson image editing [18] for seamless cloning.

4.3 Evaluation Results

To demonstrate the effectiveness of RORD, we have conducted cross-validation studies on various inpainting models by switching the training and test datasets. More specifically, existing models trained on MS-COCO [12], Places2 [29] or RORD were assessed on three datasets including PASCAL-VOC [11]. Table 2 represents that evaluation results

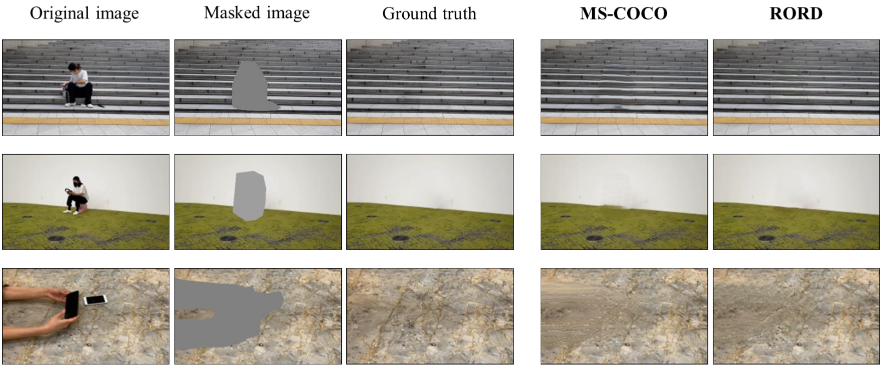


Figure 8: Visual comparison of results from the PEPSI models [20] trained with MS-COCO and RORD, respectively. The evaluation is conducted on RORD to compare the results to the object-free ground truth. Training with RORD allows the model to fill hole regions more effectively.

on PASCAL-VOC [11] and MS-COCO [12] datasets show poor performance than results on RORD. As mentioned in Section 1, the significant performance gap between the results evaluated with the conventional method or RORD is caused by the absence of ground truth background pixels behind objects. Especially, the evaluation by PASCAL and MS-COCO shows low validity to the point where there is no performance change regardless of the training dataset. Figure 7 shows several inpainting results from each dataset and its PSNR represented under the images. As can be seen in Figure 7, the image which has blur or artifacts shows rather higher PSNR when evaluated with conventional datasets. For example, in the third column, although RN results in the failure case, its PSNR value surpasses results from other methods. This tendency of the conventional datasets leads to critical drawbacks in evaluating inpainting models. In contrast, RORD has high reliability of performance measurement including appropriate ground truth images.

In this valid evaluation of object removal, training on RORD improves performance in all image inpainting models. Figure 8 shows a visual comparison of results from PEPSI model trained with MS-COCO or RORD. Indeed, the model trained on RORD synthesises more visually pleasing images than the model trained on MS-COCO. These results indicate that, as mentioned in Section 2.2, the conventional methodology for training image inpainting networks can lead to inaccurate learning, but the proposed RORD is an effective dataset for training models with large-scale and real task-specific image pairs.

5 Conclusion

In this paper, we introduced the RORD, a new large-scale object removal dataset, including paired images with and without objects along with dense annotations. RORD focuses on compensating for the absence of correct information behind the objects. Our dataset is elaborately collected to cover diverse real-world scenes and carefully annotated by experienced annotators. We demonstrate the benefits of RORD with both quantitative and qualitative performance evaluations. We expect that RORD can contribute to the field of object removal by not only providing precise ground truth for training but also serving as a benchmark for accurate performance evaluation.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000.

[3] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE Trans. Image Process.*, 12(8):882–889, 2003.

[4] Borna Besic and Abhinav Valada. Dynamic object removal and spatio-temporal rgb-d inpainting via geometry-aware adversarial learning. *IEEE Transactions on Intelligent Vehicles*, 2022.

[5] Ya-Liang Chang, Zhe Yu Liu, and Winston Hsu. Vornet: Spatio-temporally consistent video inpainting for object removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.*, 31(4):1–10, 2012.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[9] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. In *Proceedings of ACM SIGGRAPH*, pages 303–312, 2003.

[10] Selim Esedoglu and Jianhong Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88 (2):303–338, 2010.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[13] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[14] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9371–9381, 2021.

[15] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edge-connect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[16] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*, pages 394–411. Springer, 2020.

[17] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[18] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[19] Francesco Pinto, Andrea Romanoni, Matteo Matteucci, and Philip HS Torr. Seci-gan: Semantic and edge completion for dynamic objects removal. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10441–10448. IEEE, 2021.

[20] Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, and Sung-jea Ko. Pepsi: Fast image inpainting with parallel decoding network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11360–11368, 2019.

[21] Rakshith R Shetty, Mario Fritz, and Bernt Schiele. Adversarial scene editing: Automatic object removal from weak supervision. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] Yong-Goo Shin, Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Wook Kim, and Sung-Jea Ko. Pepsi++: Fast and lightweight network for image inpainting. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):252–265, 2020.

[23] Guoyao Su, Yonggang Qi, Kaiyue Pang, Jie Yang, and Yi-Zhe Song. Sketchhealer a graph-to-sequence network for recreating partial human sketches. In *Proceedings of the British Machine Vision Virtual Conference*, pages 1–14. University of Surrey, 2020.

[24] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004.

[25] Zongben Xu and Jian Sun. Image inpainting by patch propagation using patch sparsity. *IEEE Trans. Image Process.*, 19(5):1153–1165, 2010.

[26] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[27] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[28] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12733–12740, 2020.

[29] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017.

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.

[31] Wang Zhou. Image quality assessment: from error measurement to structural similarity. *IEEE transactions on image processing*, 13:600–613, 2004.