

Supplementary Material for SketchyDepth: from Scene Sketches to RGB-D Images

Anonymous ICCV submission

Paper ID 3

1. 3D photos

We wish to highlight that the supplementary material includes .mp4 files dealing with 3D photos obtained from our generated images and depth maps. These videos are best viewed within a small window of the mp4 player due to the limited resolution of the images and depth maps (128x128 pixels).

2. Additional Qualitative Results

2.1. Depth Based Creative Effects

As an extension of Fig. 5 of the main paper, in Fig. 2 we show additional examples of depth-based effects obtained with our generated images and depth maps on sketches belonging to the test set.

2.2. Colour-to-Grayscale and Transition-to-Cartoon Effects

Fig. 3 reports examples of two additional effects attainable by leveraging on the depth maps generated alongside images, namely colour-to-grayscale and transition-to-cartoon. Both effects are applied to each pixel of a generated image based on the corresponding generated depth. The colour-to-grayscale effect keeps near pixels almost unaltered while at increasing distances pixels progressively loose colour. The transition-to-cartoon instead applies a strong cartoon effect for near pixels, reducing it progressively with increasing distance values. The cartoon overlay image used for the cartoon effect is obtained starting from a generated one by a sequence of image processing operations. First of all, we transform the image to grayscale, blur it by a median filter and apply an adaptive thresholding operation to extract edges. Then we smooth the original image by applying a bilateral filter and highlight the extracted edges by colouring them in black.

2.3. Depth Based Effects: our approach vs. MiDaS

In Fig. 4, we show depth-based effects, i.e. Bokeh, light variation and fog, obtained by using either our depth maps, which are generated jointly with images, or the depth maps predicted by MiDaS [4] on the generated images. As for the Bokeh effect, we consider a fixed depth value to determine where the blur filter starts to affect the image. This parameter represents a distance threshold to separate foreground from background and it is kept fixed between the depth maps yielded by our method and MiDaS, as well as throughout all the presented examples. Then, starting at the defined distance threshold, Bokeh progressively increases image blur as the depth gets higher. All the other effects are applied using the same parameter values as in the examples shown in the main paper. Light variation decreases brightness of closer pixels with respect to those farther away. Conversely, the fog effect renders more foggy the pixels exhibiting larger depths.

Figure 4 consists of three pairs of rows. In each pair, the top row depicts the results dealing with our method, the bottom one those concerning MiDaS. Every row reports, from left to right, a generated image, the associated depth map, the Bokeh, light variation and fog effects. In the first pair of rows, we note the clear difference between our generated depth map, that looks amenable to realize effects based on discriminating between foreground, in this case a giraffe, and background, and the depth map yielded by MiDaS, that seems to fail in separating foreground from background neatly. In our results (top row), the relighting effect darkens the whole giraffe evenly, while the background is clearly brighter. In the bottom row, instead, we can see how the relighting based on the depth map computed by MiDaS fails to darken differently the different portions of the image, i.e. the giraffe and the background exhibit similar brightness. Similarly, we can see how our depth map allows



Figure 1. Left: generated 256x256 image from a test sketch. Right: small translations applied on the same sketch to improve diversity [2]

for simulating a foggy background, while the effect based on MiDaS does apply the fog to both the background as well as the giraffe. The Bokeh filter seems also rather dependent on the quality of the depth map. In the bottom row (MiDaS) the head and neck of the giraffe are blurred similarly to the background, while our depth map (top row) allows for simulating much more effectively a shallow depth of field, with the foreground object looking sharper than the background. Similar considerations can be drawn from the analysis of the third pair of rows, due to, again, the much more accurate foreground-background separation achieved by our generated depth map compared to that computed by MiDaS. Indeed, from left to right, in our results (top row) the zebra in the center of the image turns out much sharper, darker and less foggy than the background, whilst this is definitely not the case in the bottom row (MiDaS). Finally, also in the second pair of rows we can see how our depth maps are conducive to nice creative effects while those applied based on MiDaS show several issues. In fact, in the bottom row the Bokeh and fog filters are almost ineffective due to the whole image but the sky being treated as foreground. Accordingly, the light variation obscures most of the image, making, again, nearly no difference between foreground objects and the background. Overall, our experimental findings show that, most of the times, the depth map computed by MiDaS on top of an image generated from a sketch is significantly less amenable to support depth-based creative filters than the depth map we can generate jointly together with the corresponding image by the proposed network architecture and training procedure.

2.4. Image and Depth Map Generation Examples

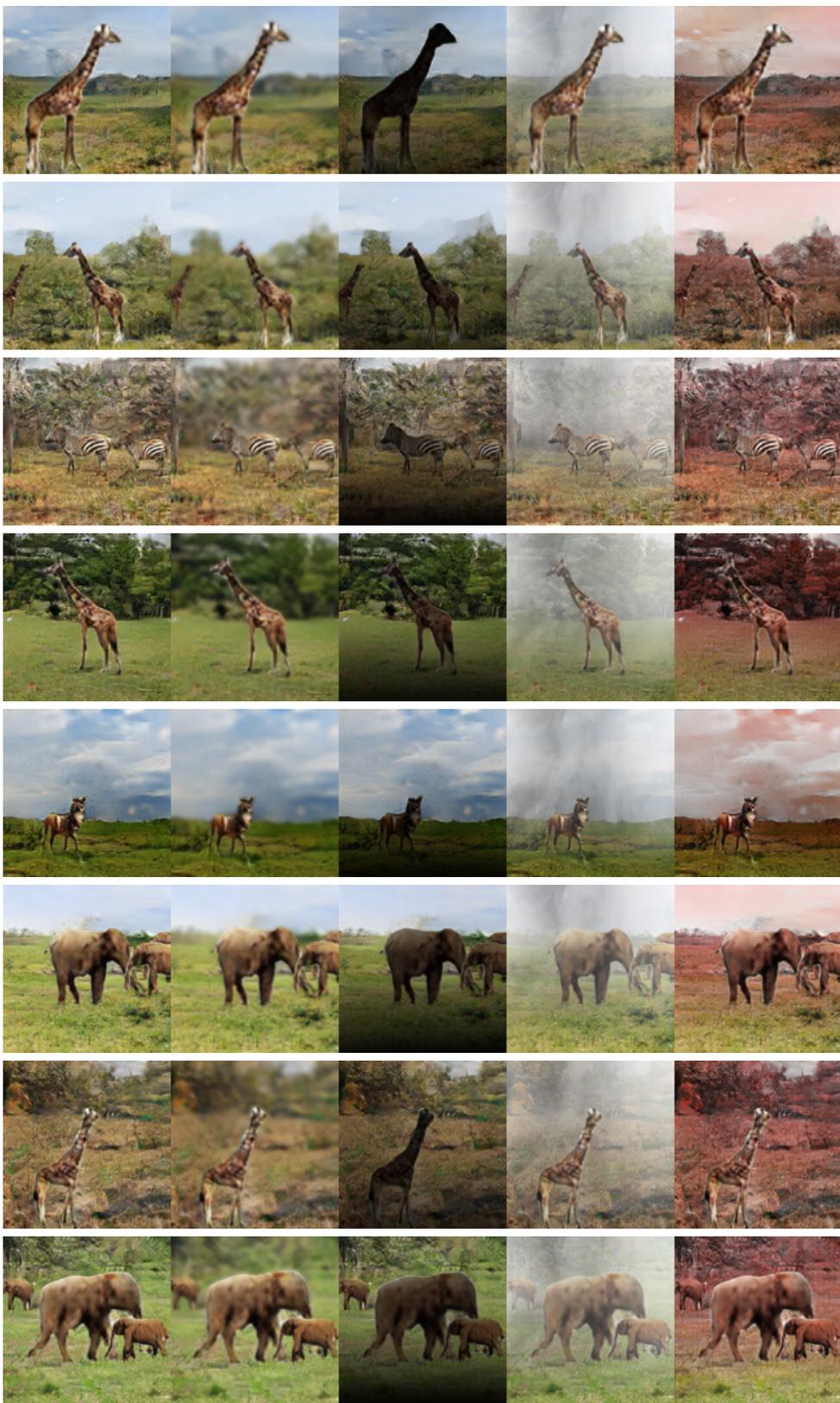
To provide a more comprehensive collection of qualitative results, in Figures 5 and 6 we report additional comparisons between our generated scene images and depth maps and the corresponding baselines.

2.5. Depth Map Sketching Examples

As an extension to figure 7 of the main paper, in figure 7 we show further examples of scene image generation through the novel depth sketching approach peculiarly enabled by our proposal.

2.6. Image Generation Resolution and Diversity.

We experimented image generation with different resolution keeping the foreground generation part fixed and training the rest of the system at 256x256 resolution. In Fig. 1 left is visible a generated test example. For what concern diversity in image generation, our work keeps dropout active also at test time, following [3]. Here we show that also "conditioning perturbation" [2], implemented as small translations on the objects of the input sketch, can be deployed in our work to increase diversity, as shown in Figure 1 right.

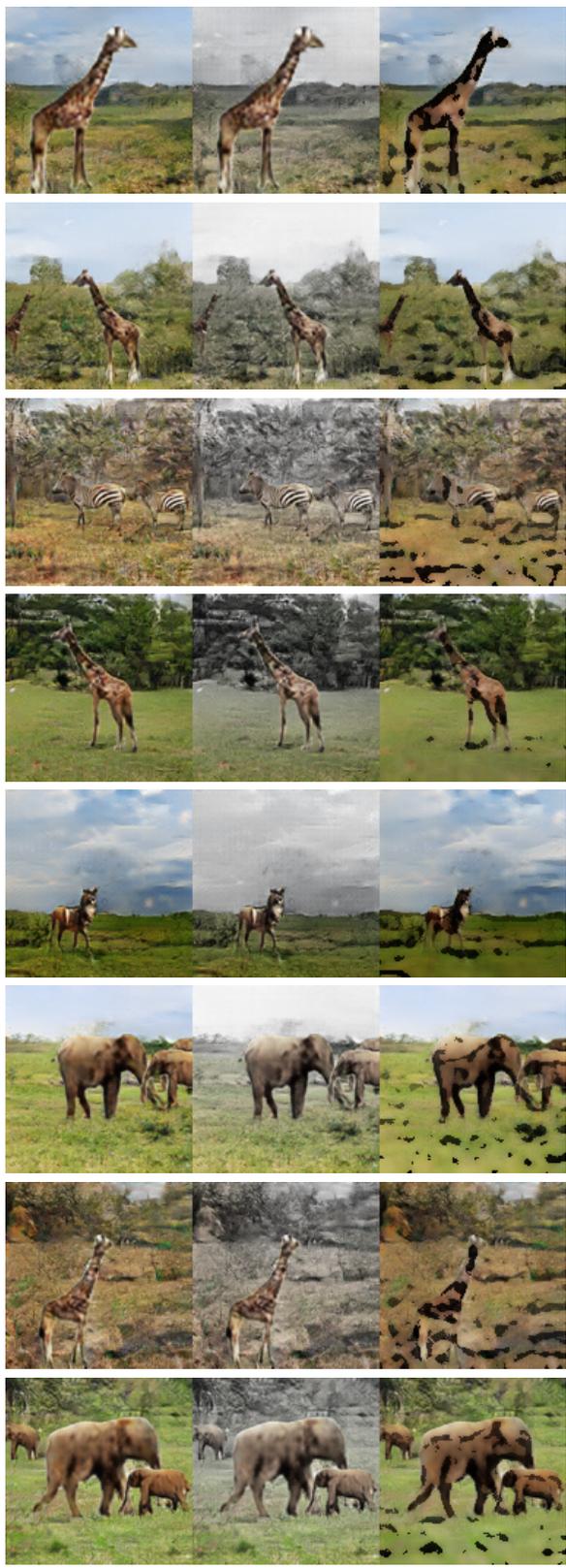


216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

Figure 2. Examples of creative effects enabled by our generated depth maps. From left to right: generated image, Bokeh, light variation, fog and hue shift.

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377



378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Figure 3. Colour-to-grayscale and transition-to-cartoon effects. From left to right: generated image, colour-to grayscale, transition-to-cartoon.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

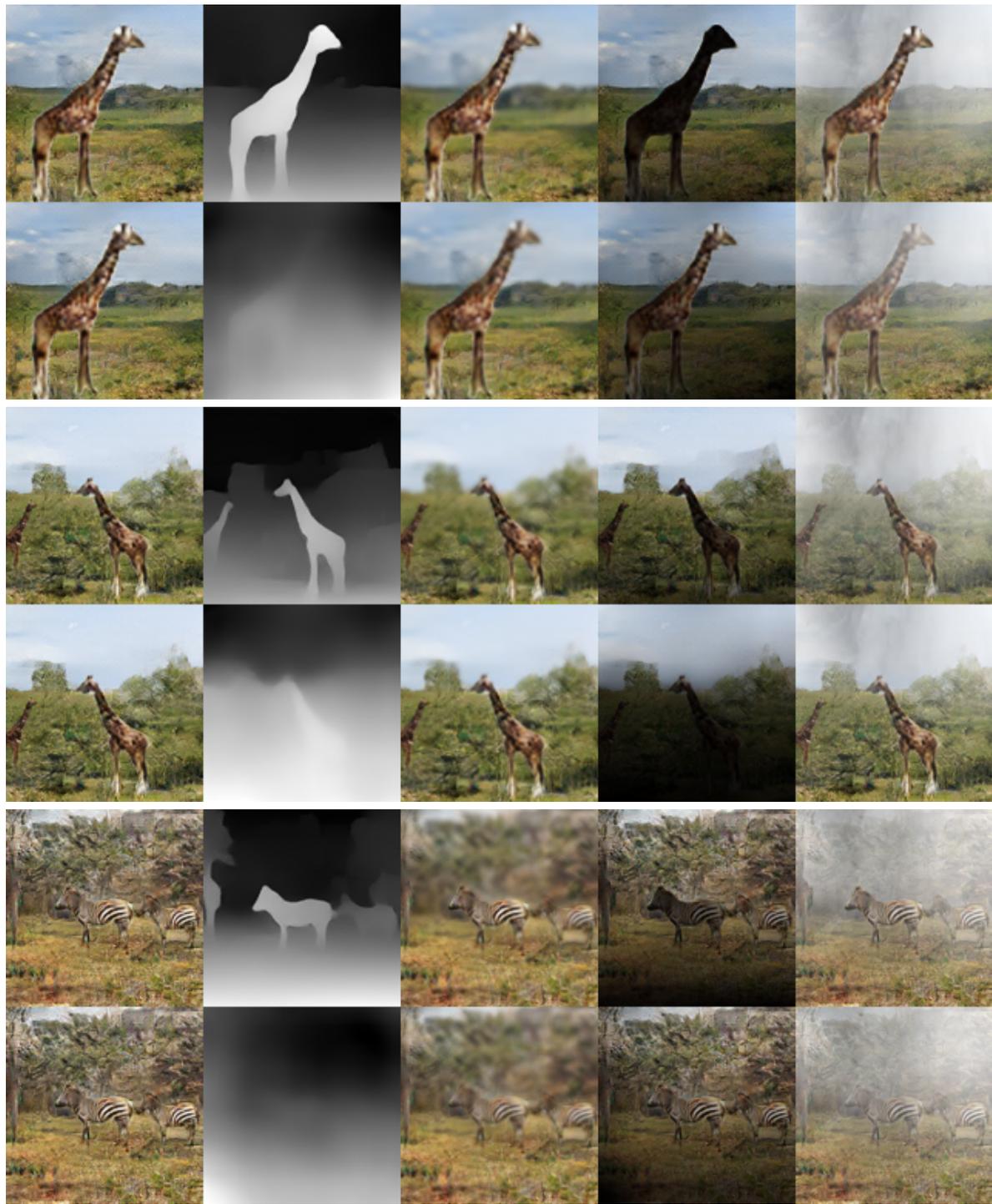


Figure 4. Comparison between depth-based effects obtained by our generated depth maps and MiDaS depth maps. Results are subdivided in three pairs of rows. In each pair, the top row deals with our depth map, the bottom one concerns MiDaS. Every row displays, from left to the right, a generated image, its depth map, the Bokeh, light variation and fog effects. See text for comments.

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647



Figure 5. Comparison between our generated images and baseline (our replication of SketchyCOCO [1]) results. In both columns, from left to the right: scene sketch, image generated by our method, image generated by the baseline method.

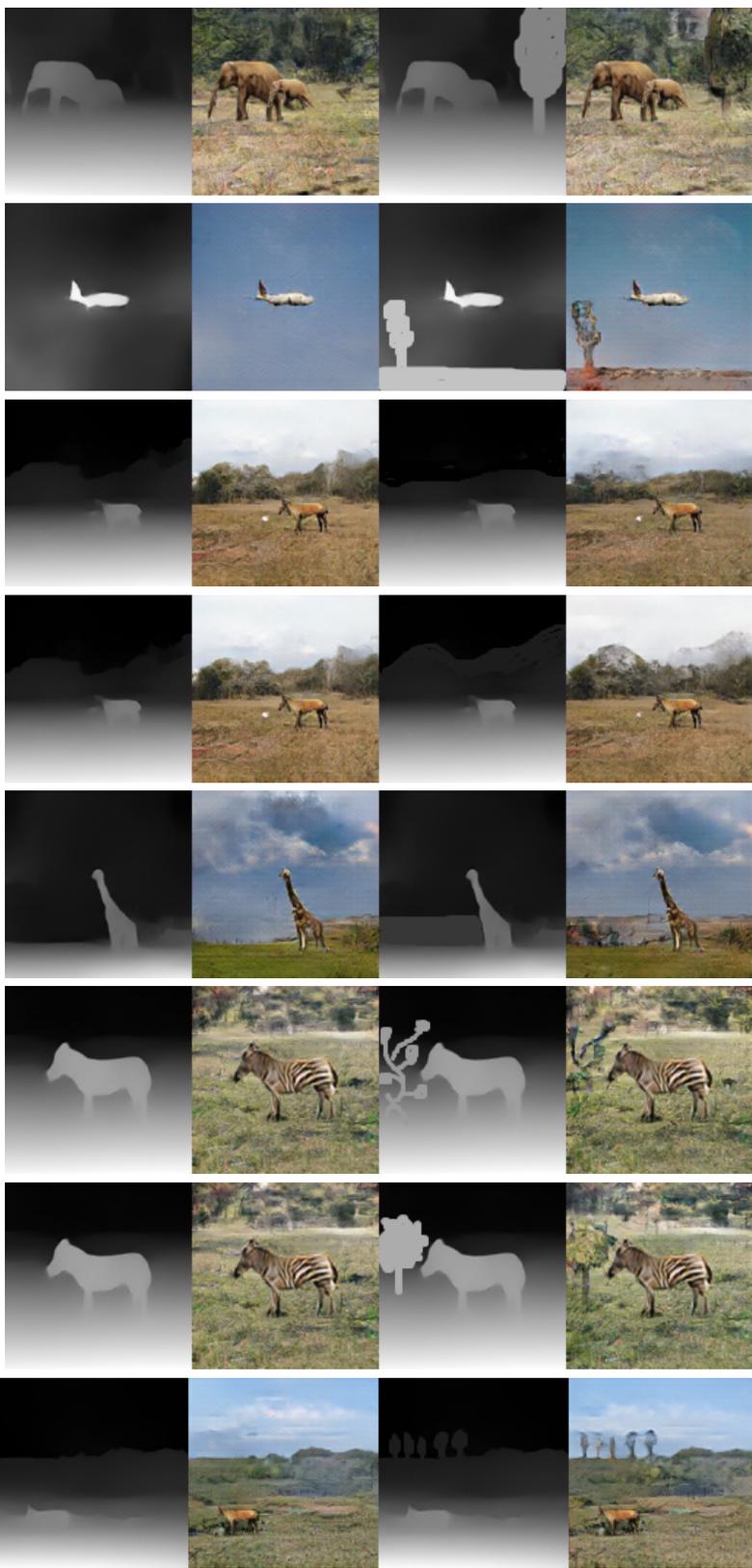
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755



Figure 6. Comparison between the depth maps generated by our method and those obtained by MiDaS. In both columns, from left to right: our generated image, our generated depth map and the depth map computed by MiDaS from our generated image.

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809



810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

Figure 7. Image generation by depth sketching. From left to right: generated depth map, generated image, sketched depth map and newly generated image.

References

[1] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5174–5183, 2020. 6

[2] Songwei Ge, Vedanuj Goswami, C Lawrence Zitnick, and Devi Parikh. Creative sketch generation. *arXiv preprint arXiv:2011.10039*, 2020. 2

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[4] Ranftl René, Lasinger Katrin, Hafner David, Schindler Konrad, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 1

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971