

000

# Pix2Surf: Learning Parametric 3D Surface 001 Models of Objects from Images

002

003

004 Anonymous ECCV submission

005

006 Paper ID 2941

007

008

## 009 S.1 Overview

010

011 In this supplementary, we provide additional details about our training (Sec-  
012 tion S.2) and inference setups (Section S.3), and details of our evaluation metrics  
013 (Section S.4). We provide an extended qualitative comparison of our method  
014 to the Image2Surf baseline (Section S.5), and for visible surface generation on  
015 real-world data (Section S.6). We show additional qualitative results for hid-  
016 den surface generation (Section S.8) and also provide more visual results for  
017 Pix2Surf (Section S.7) and more qualitative comparison to Pixel2Mesh++[3] and  
018 AtlasNet[1] (Section S.9).

019

020

## S.2 Training Details

021

022 For the **Single-View** case, we train our network in two phases. In the first  
023 phase, we train the NOCS-UV branch with a learning rate of  $1e-4$ , using the  
024 NOCS Map and the object mask as supervision. In the second phase, we add the  
025 remaining SP branch and train end-to-end until convergence, with a learning rate  
026 of  $1e-4$  for cars and  $3e-5$  for planes and chairs, and using the losses described  
027 in Section 4.1 in the paper.

028 For the **Multi-View** case, we have found that pre-training with the single-  
029 view architecture, before switching to the full multi-view architecture results in  
030 better initialization. For this purpose, we start by passing the feature  $z_m$ , directly  
031 to the SP branch without max-pooling multiple views. After pre-training, we  
032 switch to the multi-view architecture as described in Section 4.2 in the paper, by  
033 max-pooling the  $z_m$  features of all views, and concatenating both this max-pooled  
034 multi-view feature, and the single-view feature  $z_m$  for the current view as input  
035 to the MLP. We randomly pick 5 views as input during multi-view training.  
036 For our multi-view consistency loss, we need to identify corresponding pixels in  
037 different views. We sample pixels in each view as in the single-view case and find  
038 corresponding pixels based on their distance in NOCS coordinates. Two pixels  
039 are in correspondence if their NOCS distance is less than  $1e-3$ .

040 We separately train on each object category of our dataset.

041

## 042 S.3 Inference Details

043

044 One significant advantage of our explicit **continuous** parametric surface predic-  
045 tion is that we can sample the results at any resolution (e.g. points or vertices).

We generate our final predictions at a regular grid of samples in the unwrapped uv chart, obtaining a 3D location for each sample (obtained from the SP-Branch). Since we have exact correspondence to pixels of the input image, each sample also has a color value (or interpolated color value in super-resolution). Samples corresponding to background pixels are masked out. To create a mesh, we can connect neighboring foreground samples with edges. All visual results of our method in the paper are generated using this approach. We provide more details.

*Identifying foreground regions in the unwrapped chart.* Unlike AtlasNet, the shape and topology of the unwrapped surface in our chart is learned by the NOCS-UV branch, which gives the reconstructed surface more flexibility to represent arbitrary shapes and topologies. To identify foreground regions in the uv space of the unwrapped chart, we map the learned image-space foreground mask to uv space. Directly unwrap the mask by learned-uv map (two channel output from NOCS-UV branch) results in pixel cloud with holes in uv space. To solve this issue, we up-sample the image-space mask and learned-uv map from its original resolution of  $240 \times 320$  by a factor of 4 using linear interpolation before mapping mask to uv space. To avoid interpolating across  $C^0$  discontinuities of the surface, we only interpolate neighboring pixels that are mapped to similar uv locations (i.e., the gradient of their uv coordinates is below a threshold). We then map the up-sampled mask to uv space (resolution of  $128 \times 128$ ) by the up-sampled learned-uv map. Finally we up-sample the mask in uv space to the desired resolution (in paper we use  $512 \times 512$ ).

In uv space, we additionally post-process the unwrapped foreground mask by closing small holes using morphological operations. Finally, we remove outliers using the predicted 3D locations (quarried from SP-Branch) of each mask sample. A sample of the foreground mask is classified as outlier if the distance in 3D space to any of its  $m$ -nearest neighbors is larger than a threshold  $t$ . In practice, we use  $m = 6, t = 0.02$  for car,  $m = 2, t = 0.03$  for chair and  $m = 1, t = 0.02$  for airplane.

*Texturing the unwrapped chart.* Similar to the mask, directly unwrapping the image-space color values to the uv space results in a sparse set of irregular color samples in uv space. We can interpolate these samples to obtain the color value at any point in uv space by interpolating the  $k$  nearest neighbors (we use  $k = 4$  for our results).

## S.4 Evaluation Metrics

We now define the evaluation metrics used in the paper.

**A common surface representation:** Before evaluating our metrics, we convert the results of all methods to a common format to avoid biasing our results due to different surface representations. We convert all output representations to the NOCS-Map format defined in X-NOCS [2] using the ground truth camera model. The NOCS map  $\mathcal{P}$  samples the reconstructed surface from a single viewpoint, giving a point cloud where each sample has a 2D pixel coordinate

*p* and a 3D location *x*. The 3D location is defined in a canonical coordinate frame that is shared across views and across instances of the same shape category. For multi-view reconstructions, we create one NOCS-Map for each viewpoint, compute the metrics on each NOCS-Map, and average the results over all views.

The **Reconstruction Error** is measured as the 2-Way-Chamfer-Distance between the ground truth NOCS-Map  $\mathcal{P}_1$  and predicted NOCS-Map  $\mathcal{P}_2$ :

$$E_{\text{rec}} = \frac{1}{|\mathcal{P}_1|} \sum_{x_i \in \mathcal{P}_1} \min_{y_j \in \mathcal{P}_2} \|x_i - y_j\|_2^2 + \frac{1}{|\mathcal{P}_2|} \sum_{y_j \in \mathcal{P}_2} \min_{x_i \in \mathcal{P}_1} \|x_i - y_j\|_2^2.$$

The reconstruction error for hidden surfaces in Table 2 of paper is computed in the same way, but using NOCS-Maps of the hidden surfaces.

The **Correspondence Error** is measured as the squared distance between the predicted 3D location  $x_i$  and the ground truth location  $y_i$  of the same pixel:

$$E_{\text{corr}} = \frac{1}{|\mathcal{M}|} \sum_{p_i \in \mathcal{M}} \|x_i - y_i\|_2^2.$$

We only evaluate pixels  $p_i \in \mathcal{M}$  that are both in the predicted and ground truth foreground masks.

**Consistency Error** is based on the squared distance between the predicted 3D locations of corresponding pixels in different views. For each pair of views *a* and *b*, we identify corresponding pairs of pixels  $(p_i^a, p_j^b)$  as pairs having a similar ground truth 3D location in NOCS:  $\|y_i^a - y_j^b\|_2 < \epsilon$ . In practice, we set  $\epsilon = 0.001$ . We then average the squared distance between the predicted 3D locations  $x_i^a$  and  $x_j^b$  of all corresponding pixel pairs  $\mathcal{P}_{\text{corr}}^2$ :

$$E_{\text{cons}} = \frac{1}{|\mathcal{P}_{\text{corr}}^2|} \sum_{(p_i^a, p_j^b) \in \mathcal{P}_{\text{corr}}^2} \|x_i^a - x_j^b\|_2^2.$$

With the **Continuity Score**, we take a statistical approach to measure the quality of the surface continuity. We compute statistics of the  $C^0$  discontinuities in the predicted surface, and measure the similarity to the same statistics computed on the ground truth surface. The statistics are based on the 3D distance  $\|x_i - x_j\|_2$  of neighboring foreground pixels  $p_i$  and  $p_j$ . Pixels with a large difference are likely to lie on the border of a  $C^0$  discontinuity of the predicted surface. We compute a histogram  $h$  of this 3D distance over all neighboring pixels:

$$h_i = |\{(p_i, p_j) \in \mathcal{P}_{\text{neighbors}}^2 \mid t_i \leq \|x_i - x_j\|_2 < t_{i+1}\}|,$$

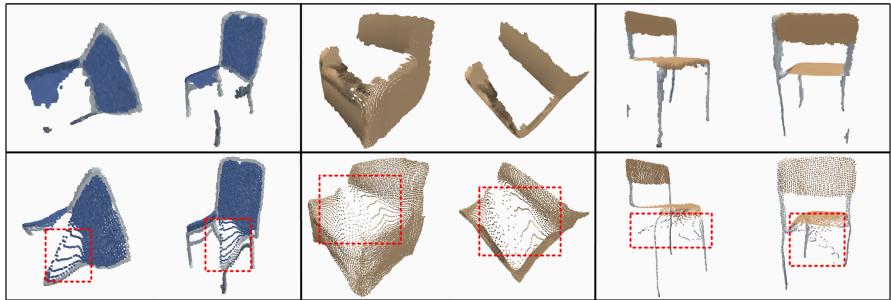
where  $t_i$  are the boundaries of the histogram bins and  $\mathcal{P}_{\text{neighbors}}^2$  is the set of all neighboring pixel pairs. We use a 4-connected neighborhood and choose 20 bins with bin edges spaced uniformly in  $[0.05, \sqrt{3}]$ . We measure the similarity of two histograms as the correlation of their normalized bins:

$$S_{\text{cont}} = \frac{1}{\sum_k h_k \sum_j h_j^{\text{gt}}} \sum_i h_i h_i^{\text{gt}}$$

Note that unlike the other errors we use as evaluation metrics, this is defined as a score, where higher values imply better continuity of the reconstructed surface.

## S.5 Qualitative Comparison to Image2Surf

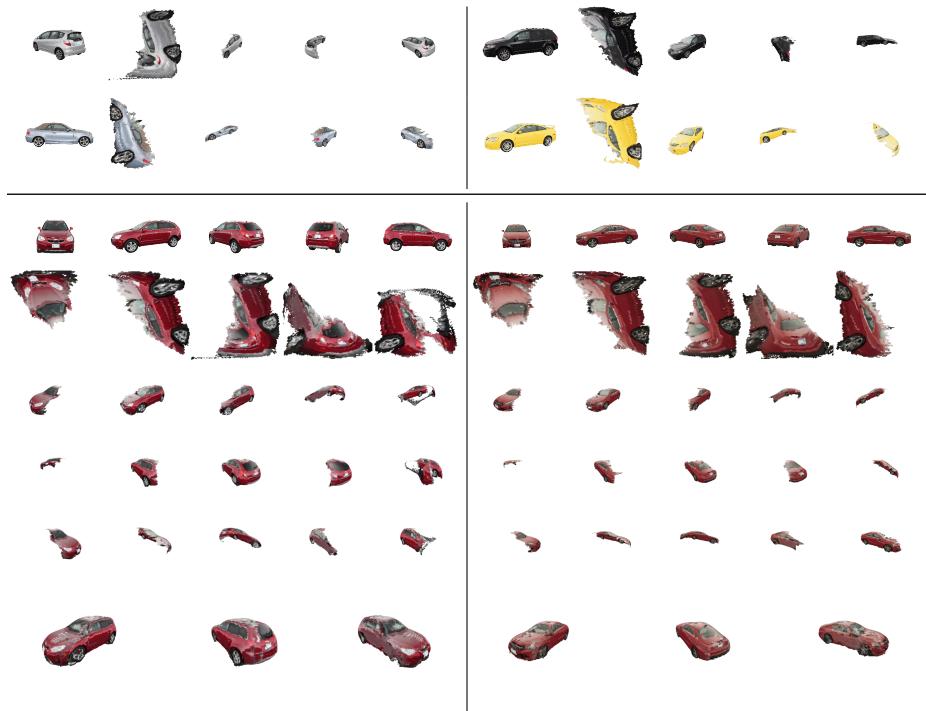
We show more qualitative comparisons between our baseline Image2Surf and Pix2Surf in paper Sec 5.1. Image2Surf has a fatal problem to make “cut” around the occlusion boundary (i.e., wrong  $C^0$  discontinuities), which is reflected both in the red rectangle in Figure S1 and continuity score in Table 1 in paper.



**Fig. S1.** Qualitative Comparison to Image2Surf. The first row are the results of Pix2Surf and the second row are for Image2Surf. Each instance is viewed from 2 different viewpoints. Image2Surf wrongly connects disjoint parts and results in strong distortions, which are solved by Pix2Surf’s learned chart.

## 180 S.6 Qualitative Results on Real-World Data

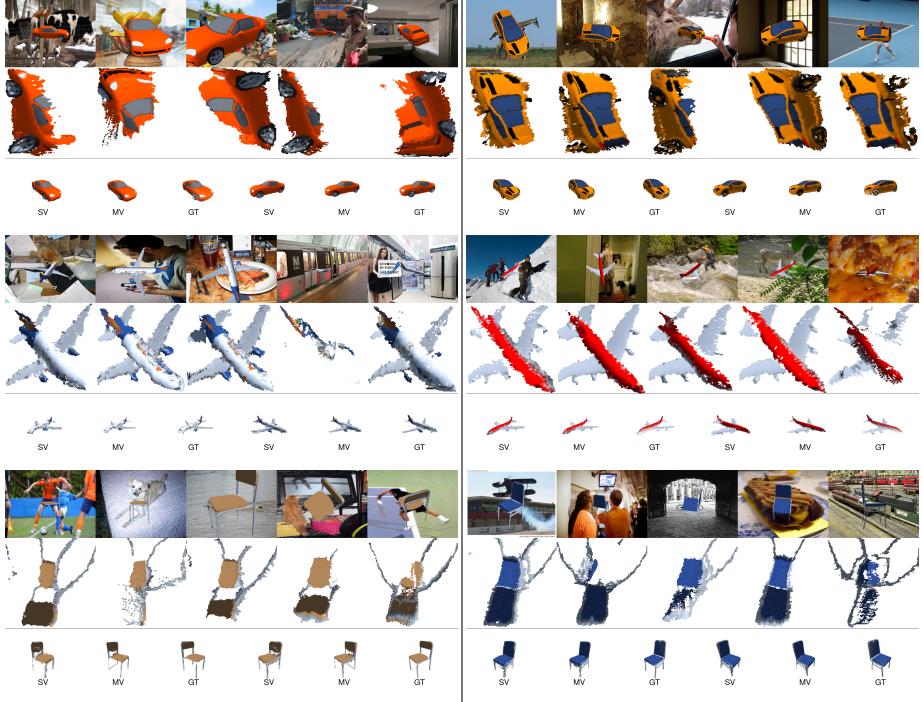
181 We show more results for generalization to real world data mentioned in Sec 5.3  
 182 in paper. In Figure S2, we show single- and multi-view results for Pix2Surf that  
 183 is trained on ShapeNet COCO and inference on real world car. Note that the  
 184 texture in each view separately is better than the multi-view aggregation. This is  
 185 caused by the different light condition from different viewpoints. As our main  
 186 concern in this paper is not to fuse the texture from multiple views, we leave the  
 187 improvement of the texture to future works.



212 **Fig. S2.** Real world image generalization. The top part is single-view visualization:  
 213 input image, unwrapped chart with texture and 3 viewpoints of the reconstruction  
 214 for each instance. The bottom part is multi-view aggregation visualization. For every  
 215 instance, each row is: input images, unwrapped charts with texture, 3 viewpoints for  
 216 each view's result separately and finally multi view aggregation.

## 225 S.7 More Results

226 Figure S3 shows more results of Pix2Surf including the learned UV map (as  
 227 shown in Figure 4) and reconstruction outputs of both single-view and multi-view  
 228 architectures. See the caption for the details.



254 **Fig. S3.** Single-view and multi-view Pix2Surf reconstruction results. The results for each  
 255 object are presented in three rows. The first row shows five input views. Note that we do  
 256 not have camera parameters for any of these views. The second row shows the per-view  
 257 UV space that is generated by the multi-view variant of Pix2Surf. The UV space is  
 258 not directly constrained by any loss; the flattening of the objects that we can observe  
 259 and the large degree of consistency between different views is an emergent property of  
 260 our network. In the third row, we show, from left to right, (a) the reconstructed 3D  
 261 surface obtained by merging Pix2Surf single-view reconstructions (SV), (b) the Pix2Surf  
 262 multi-view reconstruction (MV), and (c) the ground truth reconstruction (GT). The  
 263 last three columns show the same results from a different viewpoint. Note the reduction  
 264 in the number of gaps and surface discontinuities when comparing the multi-view to  
 265 the single-view results.

## S.8 Qualitative Results for Hidden Surface Generation

The following table provides more visual results of Pix2Surf Two-Intersection version (Sec 5.2 in paper), and comparison with X-NOCS [2]. Pix2Surf can easily be extended to capture the invisible surface and is more accurate and smooth than X-NOCS.

	View 1			View 2			View 3		
	Pix2Surf (sv)	X-NOCS (sv)	Ground Truth	Pix2Surf (sv)	X-NOCS (sv)	Ground Truth	Pix2Surf (sv)	X-NOCS (sv)	Ground Truth
280									
281									
282									
283									
284									
285									
286									
287									
288									
289									
290									
291									
292									
293									
294									
295									
296									
297									
298									
299									
300									

### 315 S.9 Qualitative Comparisons

316 As shown in the Table 4 in the paper, the following table demonstrates qualitative  
 317 comparisons among our Pix2Surf (both single-view and multi-view architectures),  
 318 AtlasNet[1], and Pixel2Mesh++ [3]. The colors in AtlasNet results show different  
 319 output patches.  
 320

	View 1			View 2			View 1	
	Single View	Multi-View	Ground Truth	Single View	Multi-View	Ground Truth	Atlas Net	Pixel2 Mesh++
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								

View 1			View 2			View 1	
Single View	Multi-View	Ground Truth	Single View	Multi-View	Ground Truth	Atlas Net	Pixel2 Mesh++
		<img alt="Ground truth of a brown					

## 405 References

- 406
- 407 1. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Atlasnet: A papier-  
408 mâché approach to learning 3d surface generation. In: Proc. of CVPR (2018)
- 409 2. Sridhar, S., Rempe, D., Valentin, J., Bouaziz, S., Guibas, L.J.: Multiview aggregation  
410 for learning category-specific shape reconstruction. In: Proc. of NeurIPS (2019)
- 411 3. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2mesh++: Multi-view 3d mesh generation  
412 via deformation. In: Proc. of ICCV (2019)
- 413
- 414
- 415
- 416
- 417
- 418
- 419
- 420
- 421
- 422
- 423
- 424
- 425
- 426
- 427
- 428
- 429
- 430
- 431
- 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
- 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
- 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
- 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
- 864
- 865
- 866
- 867
- 868
- 869
- 870
- 871
- 872
- 873
- 874
- 875
- 876
- 877
- 878
- 879
- 880
- 881
- 882
- 883
- 884
- 885
- 886
- 887
- 888
- 889
- 890
- 891
- 892
- 893
- 894
- 895
- 896
- 897
- 898
- 899
- 900
- 901
- 902
- 903
- 904
- 905
- 906
- 907
- 908
- 909
- 910
- 911
- 912
- 913
- 914
- 915
- 916
- 917
- 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
- 972
- 973
- 974
- 975
- 976
- 977
- 978
- 979
- 980
- 981
- 982
- 983
- 984
- 985
- 986
- 987
- 988
- 989
- 990
- 991
- 992
- 993
- 994
- 995
- 996
- 997
- 998
- 999
- 1000