International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 311-317 www.ijaem.net ISSN: 2395-5252

GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis

[1,2]student, Davda Brinda [3]student,Jadeja Varshaba.
Dr.Darshana Patel,Dishita Mashru,,Aditiba Jadeja
Corresponding Author:Soniya Aghera

Date of Submission: 11-03-2024 Date of Acceptance: 21-03-2024

ABSTRACT: Collision-avoidance is a crucial research topic in robotics. Designing a collision-avoidance algorithm is still a challenging and open task, because of the requirements for navigating in unstructuredand dynamic environments using limited payload and computing resources on board micro aerial vehicles. This article presents a novel depth-based collision-avoidance method for aerial robots, enabling high-speed flights in dynamic environments. First of all, a depth-based Euclidean distance field mapping algorithm is generated.Then, the proposed Euclidean distance field mapping strategy is integrated with a rapid-exploration random tree to construct a collision-avoidance system. The experimental results show that the proposed collision-avoidance algorithm has a robust performance at high flight speeds in challenging dynamic environments. The experimental results show that the proposed collision-avoidance algorithm can perform faster collision-avoidance maneuvers when compared to the state-of-art algorithms (the average computing time of the collision maneuver is 25.4 ms, while the minimum computing time is 10.4 ms). The average computing time is six times faster than one baseline algorithm. Additionally, fully autonomous flight experiments are also conducted for validating the presented collision-avoidance approach.

KEYWORDS: Micro aerial vehicles; collision-avoidance; distance field; depth sensor

I. INTRODUCTION

Convolution Neural Networks (CNN s), have shown great potential in image analysis. One problem with training machine learning models on images is that the use of datasets that are small and lack of diversity ends up in ineffective and inaccurate outcomes.

As reasoning in 3D image is the key for applications in robotics, virtual reality or data augmentation, many recent works consider the task of 3D-aware image synthesis, aiming at photorealistic image generation with specific control over the camera pose. In contrast to 2D generative adversarial networks, approaches for 3D-aware image synthesis learn a 3D scene representation that is explicitly mapped to an image using differentiable rendering techniques, thus providing control over each scene content and viewpoint. Since 3D supervision or posed images are often hard to obtain in practice, recent works try and solve this task using 2D supervision only. Towards this goal, existing approaches generate discretized 3D representations, i.e., a voxel-grid representing either the total 3D object or intermediate 3D features. While modeling the 3D object in color space allows for exploiting differentiable rendering, the cubic memory growth of voxel-based representations limits to low resolution and leads to visible artifacts[14]. Intermediate 3D features are more compact and scale better with image resolution.

Voxel-based approaches for 3D-aware image synthesis either generate a voxelized 3D model (e.g., PlatonicGAN[6]) or learn associated abstract 3D feature representation (e.g., HoloGAN). This ends up in discretization artifacts or degrades view-consistency of the generated images because of the learned neural projection function.

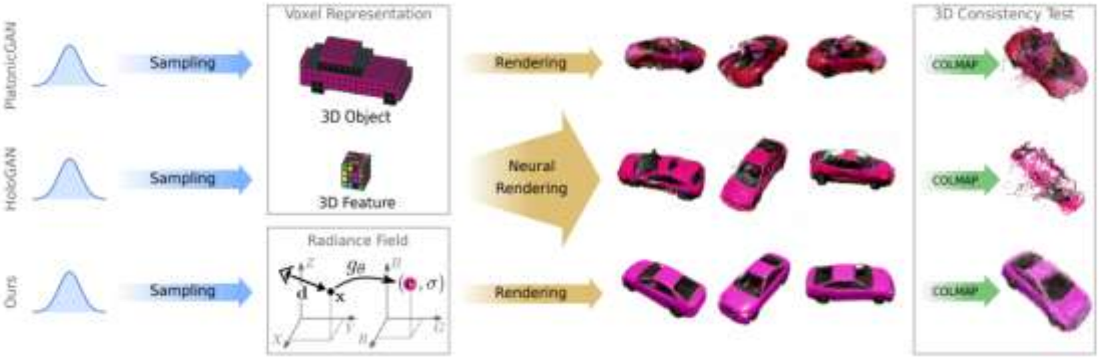


Fig: 1 Rendering a 2D image into 3D image

In this paper, a generative model for neural radiance fields (bottom) is introduced, which represents the scene as a continuous function ‘gθ’ that maps a location ‘x’ and also viewing direction ‘d’ to a colour value ‘c’ and a volume density ‘σ’. This model permits for generating 3D consistent images at high spatial resolution. It visualizes 3D consistency by running a multi-view stereo algorithm (COLMAP) on several outputs of each method (right). It is important to note that all models have been trained using 2D supervision only (i.e., from unposed RGB images).In this paper, a trial to demonstrate that the dilemma between coarse outputs and entangled latent will be resolved using conditional radiance fields has been made, a conditional variant of a recently planned continuous representation for novel view synthesis. Specifically, this proposed method creates the following contributions:

i)Propose GRAF, a generative model for radiance fields for high-resolution 3D-aware image synthesis from unposed images.

ii)Introduction to a patch-based discriminator that samples the image at multiple scales and that is key to find out high-resolution generative radiance fields efficiently.

iii)Evaluate the proposed approach on synthetic and real datasets.

This approach compares favourably to state-of-the-art methods in terms of visual fidelity [3] and 3D consistency whereas generalizing to high spatial resolutions.

II. RELATED WORK

3D Aware image synthesis: Learning-based novel view synthesis has been intensively investigated within the literature. These methods generate unseen views from the same object and usually need camera viewpoints as supervision. Whereas recent works generalize across different objects without requiring to train an individual network per object, they do not yield a full probabilistic generative model for drawing unconditional random samples. Previously proposed methods require 3D supervision or assume 3D information as input[10]. E.g. Texture Fields synthesize novel textures conditioned on a particular 3D shapeThe proposed model learns a generative model for each shape and texture from 2D images alone. Some of the defined models like PlatonicGAN learns a textured 3D voxel representation from 2D images using differentiable rendering techniques[4]. However, such voxel-based representations are memory intensive, precluding image synthesis at high image resolutions.

In this paper, the proposed model tries to avoid those memory limitations by using a continuous representation that permits for rendering images at arbitrary resolution. HoloGAN [6] and some related works learn a low-dimensional 3D feature combined with a learnable 3D-to-2D projection. However, as proven by random experiments, learned projections will result in entangled latent (e.g., object identity and viewpoint), particularly at high resolutions. Whereas 3D consistency can be encouraged using additional constraints, we take advantage of differentiable volume rendering techniques which do not need to be learned and thus incorporate 3D consistency into the generative model by design.

Implicit Representations: Recently, implicit representations of 3D geometry have gained quality in learning-based 3D reconstruction. Key advantages over voxel or mesh-based methods are that they do not discretize space and are not restriced in topology. Recent hybrid continuous grid representations extend implicit representations to complicated or large scale scenes but require 3D input and do not consider texture. Another line of works proposes to learn continuous shape and texture representations from posed multi-view images only, by making the rendering process differentiable. As these models are limited to single objects or scenes of little geometric complexity propose to represent scenes as neural radiance fields that allow for multi-view consistent novel-view synthesis of more complicated, real-world scenes from posed 2D images. They demonstrate compelling results on this task, however, their method needs several posed views, needs to be retrained for every scene, and can't generate novel scenes

III. METHOD

This proposed work considers the matter of 3D-aware image synthesis, i.e., the task of generating high-fidelity images whereas providing explicit control over camera rotation and translation. It is a point of discussion to represent a scene by its radiance field as such a continuous representation scales well with respect to image resolution and memory consumption whereas allowing a physically-based and parameter-free projective mapping within the following, the model was initialized in brief review of Neural Radiance Fields (NeRF)[1] that forms the idea for the proposed Generative Radiance Field (GRAF) model.

3.1 Neural Radiance Field:

Neural Radiance Fields NeRF or higher referred to as Neural Radiance Fields [2] is a state-of-the-art method that generates novel views of complicated scenes by optimizing an underlying continuous volumetric scene function using a sparse set of input views. The input will be provided as a blender model or a static set of images.

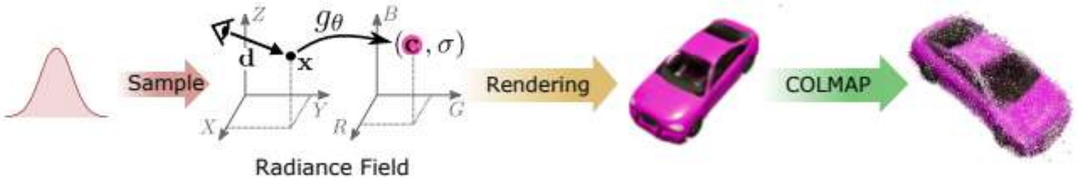
3.1.1 Positional encoding

A radiance field could be a continuous mapping from a 3D location and a 2D viewing direction to an RGB colour value. It first maps a 3D location x ∈ R3 and a viewing direction d ∈ S2 to a higher-dimensional feature representation using a fixed positional encoding that is applied element-wise to all three elements of x and d:

$$\gamma(p) = (\sin(20\pi p), \cos(20\pi p), \sin(21\pi p), \cos(21\pi p), \sin(22\pi p), \cos(22\pi p), \ldots) \qquad (1)$$

Above figure shows the proposed model. The generator Gθ takes camera matrix K, camera pose ξ, 2D sampling pattern ν and shape/appearance codes zs ∈ R m/za ∈ R n as input and predicts an image patch P'. The discriminator Dφ compares the synthesized patch P' to a patch P extracted from a real image I.

At inference time, it predicts one colour value for every image pixel. However, at training time, this is too expensive. Therefore, the model instead predict a fixed patch of size K × K pixels which is randomly scaled and rotated to provide gradients for the entire radiancefield.



3.2.1Generator

The model samples the camera matrix K, the intrinsic parameters, the camera pose psi(ξ) which are the extrinsic parameter and a patch sampling pattern ν.

ν = (u, s) determines the center u = (u, v) ∈ R2 and scale s ∈ R + of the virtual K × K patch P(u, s) which is aimed to generate. This permits the model to use a convolutional discriminator independent of the image resolution. It arbitrarily draws the patch centre u ∼ U(Ω) from a uniform distribution over the image domain Ω and the patch scale s from a uniform distribution s ∼ U([1, S]) where S = min(W, H)/K with W and H denoting the width and height of the target image.

Moreover, it makes sure that the entire patch is within the image domain Ω. The shape and appearance variables zs and za are drawn from shape and appearance distributions zs ∼ps and za ∼ pa, severally. In random experiments the model uses a standard Gaussian distribution for both ps and pa [10].

Ray Sampling:

$$P(u, s) = \left\{ (sx + u, sy + v) \;\middle|\; x, y \in \left\{ -\frac{K}{2}, \ldots, \frac{K}{2} - 1 \right\} \right\} \qquad (5)$$

The K × K patch P(u, s) is determined by a set of 2D image coordinates which describe the location of every pixel of the patch in the image domain Ω. Note that these coordinates are real numbers, not discrete integers that permits us to continuously evaluate the radiance field. The corresponding 3D rays are uniquely determined by P(u, s), the camera pose ξ and the intrinsic K. Here, the model denotes the pixel/ray index by r, the normalized 3D rays by dr and the number of rays by R where R = K2 during training and R = W H during inference.

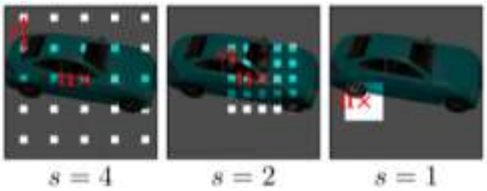
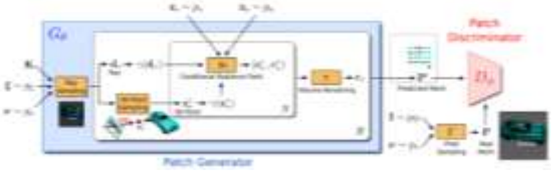


Fig. 3: Ray Sampling on images

3D Point Sampling: To approximate the intractable volumetric projection integral propose a stratified sampling approach that allows to query the network at continuous intervals instead of a discretized grid.[12]



Conditional Radiance Field: The radiance field is denoted by a deep fully-connected neural network with parameters θ that maps the positional encoding (cf. Eq. (1)) of 3D location x ∈ R^3 and viewing direction d ∈ S^2 to an RGB colour value c and a volume density σ:

$$g_\theta : R^{Lx} \times R^{Ld} \times R^{Ms} \times R^{Ma} \to R^3 \times R^+ \quad (\gamma(x), \gamma(d), \mathbf{z}_{s},$$

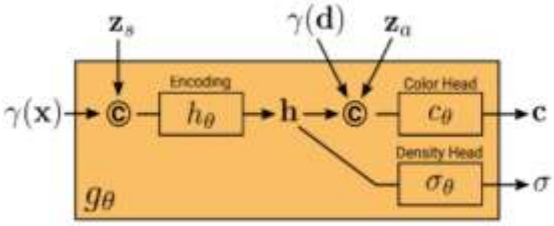


Fig. 5: Conditional Radiance Field

The network architecture of the proposed conditional radiance field gθ is illustrated above.

It first computes a shape encoding h from the positional encoding of x and the shape code zs. A density head σθ transforms this encoding to the volume density σ. For predicting the colorc at 3D location x, it concatenates h with the positional encoding of d and the appearance code za and pass the resulting vector to a color head cθ. The model computes σ independently of the viewpoint d and the appearance code za to encourage multi-view consistency whereas disentangling shape from appearance. This encourages the network to use the latent codes zs and za to model shape and appearance, respectively, and allows for manipulating them separately during inference.[9] More formally:

$$h\theta : R^{Lx} \times R^{Ms} \to R^{H} \qquad (\gamma(x), zs)\ 7\!\to h \quad (7)$$

$$c\theta : R^{H} \times R^{Ld} \times R^{Ma} \to R^{3}$$
$$(h(x, z_s), \gamma(d), z_a)\ 7\!\to c \qquad (8) \backslash$$
$$\sigma\theta : R^{H} \to R + h(x, z_s)\ 7\!\to \sigma \qquad (9)$$

All mappings (hθ, cθ and σθ) are implemented using fully connected networks with ReLU activations. To avoid notation clutter, the equation uses the same symbol θ to denote the parameters of each network.

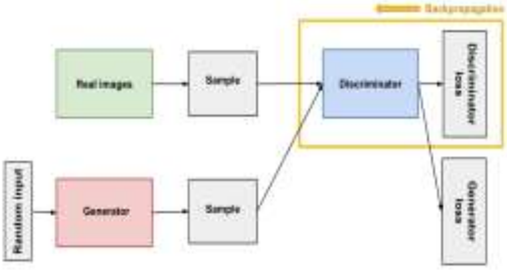


Fig. 6: Discriminator Model

The discriminator classifies each real data and fake data from the generator. The discriminator's training data comes from two sources first is Real data instances, such as real pictures of people. The discriminator uses these instances as positive examples during training and also the second is Fake data instances created by the generator. The discriminator uses these instances as negative examples throughout training.[8]

Here, The discriminator Dφ is implemented as a convolutional neural network which compares the predicted patch P' to a patch P extracted from a real image I drawn from the data distribution pD. For extracting a K × K patch from real image I, first draw ν = (u, s) from the same distribution pν which will be used for drawing the generator patch above. The model then samples the real patch P by querying I at the 2D image coordinates P(u, s) using bilinear interpolation. In the following, the model uses Γ(I, ν) to denote this bilinear sampling operation. Note that this discriminator is similar to PatchGAN, except that it allows for continuous displacements u and scales s while PatchGAN uses s = 1. It's more necessary to notice that it does not downsample the real image I based on s, but instead query I at sparse locations to retain high-frequency details.

IV. EXPERIMENT

Datasets: The model considered two synthetic datasets in our experiments. To analyze the approach in a controlled setting, the model rendered 150k Chair images from Photoshapes[16,19]. Also, the Carla Driving simulator[15] was used to create 10k images of 18 car models with randomly sampled colors and realistic texture and reflectance properties (Cars). The model validated the proposed approach on three real-world datasets. It used the Faces dataset which comprises celebA and celebA-HQ for imagesynthesis up to resolution 128^2 and 512^2 pixels, respectively. In addition, the Cats dataset and the Caltech-UCSD Birds-200-2011 dataset was considered.

Baselines: The proposed approach was compared to two state-of-the-art models for 3D-aware image synthesis using the authors' implementations [12,13]: PlatonicGAN generates a voxel-grid of the 3D object which is projected to the image plane using differentiable volumetric rendering. HoloGAN[6] instead generates an abstract voxelized feature representation and learns the mapping from 3D to 2D using a combination of 3D and 2D convolutions. To analyze the consequences of a learned projection this approach considers a modified version of HoloGAN (HoloGAN w/o 3D Conv) in which it reduces the capacity of the learned mapping by removing the 3D convolutional layers. For reference, the proposed model also compared the results to a state-of-the-art 2D GAN model with a ResNet architecture[6,20].

PGAN
HGAN
Ours
PGAN
HGAN
Ours

The following table describes the comparison of Camera Pose Interpolations for Cars and Chairs at image resolution 642 pixels for PlatonicGAN, HoloGAN and the proposed Approach[17].

	Chairs	Birds	Cars	Cats	Faces
2D- GAN	59	24	66	18	15
PlatonicGAN	199	179	169	318	321
HoloGAN	59	78	134	27	25
Proposed Model (Ours)	34	47	30	26	25

Table 1: FID at Image resolution 64^2 pixels

This approach quantifies image fidelity using the Frechet Inception Distance (FID) and additionally report the Kernel Inception Distance (KID). To assess 3D consistency, the approach performs 3D reconstruction for images of size 2562 pixels using COLMAP. The proposed approach adopts Minimum Matching Distance (MMD) to measure the chamfer distance (CD) between 100 reconstructed shapes and their closest shapes in the ground truth for quantitative comparison and showqualitative results for the reconstructions.

Compare generative radiance field with voxel-based approaches:

The proposed model is first compared against the baselines using an image resolution of 642 pixels. All of the methods are able to disentangle object identity and camera viewpoint.

However, PlatonicGAN has difficulties in representing thin structures and both PlatonicGAN and HoloGAN lead to visible artifacts in comparison to the proposed model. This is also reflected by larger FID scores in Table 1. On Faces and Cats, HoloGAN achieves FID scores same as our approach as each datasets exhibit only little variation in the azimuth angle of the camera whereas the other datasets cover larger viewpoint variations[18]. This implies that it's more harder for HoloGAN to accurately capture the appearance of objects from different viewpoints due to its low-dimensional 3D feature representation and also the learnable projection. In contrast, the continuous representation of the proposed approach doesn't need a learned projection and renders high-fidelity images from arbitrary views.

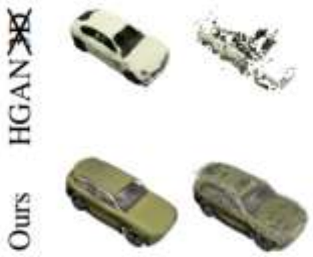
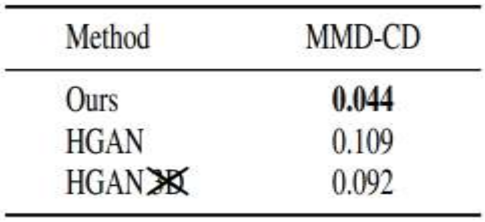


Fig. 8:3D Reconstuction



Method	MMD-CD
Ours	**0.044**
HGAN	0.109
HGAN~~3D~~	0.092

Table 2: Reconstuction Accuracy

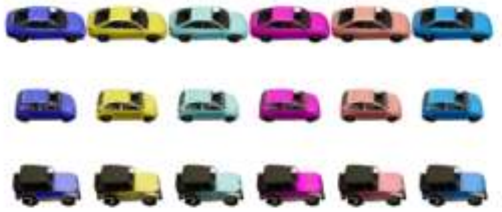


Fig.9:Disentangling Shape /Appearance[7]

V. CONCLUSION

Generative Radiance Fields (GRAF) has been utilized for high-resolution 3D-aware image synthesis for reducing the reconstruction disentanglement and improving the accuracy of rendering the 2D image to a 3D image, as compared to PlatonicGAN and HoloGAN. The proposed method shows that this framework is able to generate high resolution images with better multi-view consistency compared to voxel-based approaches. However, the results are limited to simple scenes with single objects. It has been observed that incorporating inductive biases, e.g., depth maps or symmetry, will allow for extending the model to even more challenging real-world scenarios in the future.

REFERENCES

[1]. B. Mildenhall, P. P. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. Proc. of the European Conf. on Computer Vision (ECCV).

[2]. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Proc. of the European Conf. on Computer Vision (ECCV), 2020.

[3]. A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In Proc. of the International Conf. on Learning Representations (ICLR), 2019.

[4]. A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. arXiv.org, 1608.04236, 2016.

[5]. P. Henzler, N. J. Mitra, and T. Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In Proc. of the IEEE International Conf. on Computer Vision (ICCV), 2019.

[6]. T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In Proc. of the IEEE International Conf. on Computer Vision (ICCV), 2019.

[7]. J. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman. Visual object networks: Image generation with disentangled 3d representations. In Advances in Neural Information Processing Systems (NeurIPS), 2018.

[8]. F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of gan-generated fake images over social networks. In Proc. IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR), 2018.

[9]. N. L. Max. Optical models for direct volume rendering. IEEE Transactions on Visualization and Computer Graphic (TVCG), 1(2):99–108, 1995.

[10]. J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[11]. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[12]. P. Henzler, N. J. Mitra, and T. Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In Proc. of the IEEE

International Conf. on Computer Vision (ICCV), 2019.

[13]. A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In Proc. of the International Conf. on Learning Representations (ICLR), 2019.

[14]. A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. arXiv.org, 1608.04236, 2016.

[15]. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In Proc. Conf. on Robot Learning (CoRL), 2017.

[16]. K. Park, K. Rematas, A. Farhadi, and S. M. Seitz. Photoshape: Photorealistic materials for large-scale shape collections. Communications of the ACM, 2018.

[17]. W. Zhang, J. Sun, and X. Tang. Cat head detection - how to effectively exploit shape and texture features. In Proc. of the European Conf. on Computer Vision (ECCV), 2008.

[18]. W. Nie, T. Karras, A. Garg, S. Debhath, A. Patney, A. B. Patel, and A. Anandkumar. Semisupervisedstylegan for disentanglement learning. In Proc. of the International Conf. on Machine learning (ICML), 2020.

[19]. A. Noguchi and T. Harada. RGBD-GAN: unsupervised 3d representation learning from natural image datasets via RGBD image synthesis. In Proc. of the International Conf. on Learning Representations (ICLR), 2020.

[20]. G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.