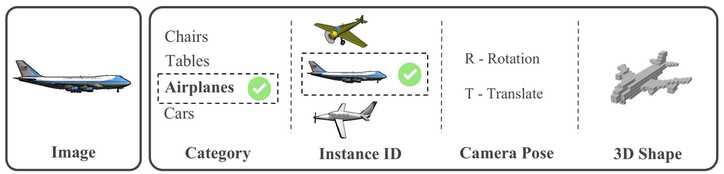
Learning Single-View 3D Reconstruction with Limited Pose Supervisions

Guandao Yang[1], Yin Cui[1,2], Serge Belongie[1,2], Bharath Hariharan[1]

[1] Department of Computer Science, Cornell University [2] Cornell Tech

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Problem Setup

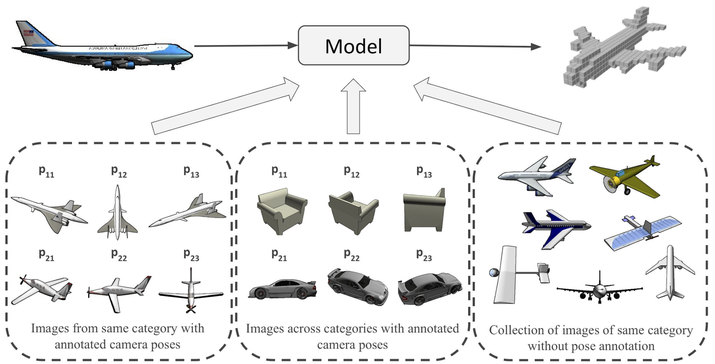


Cheap Plentiful ⟷ Expensive Limited

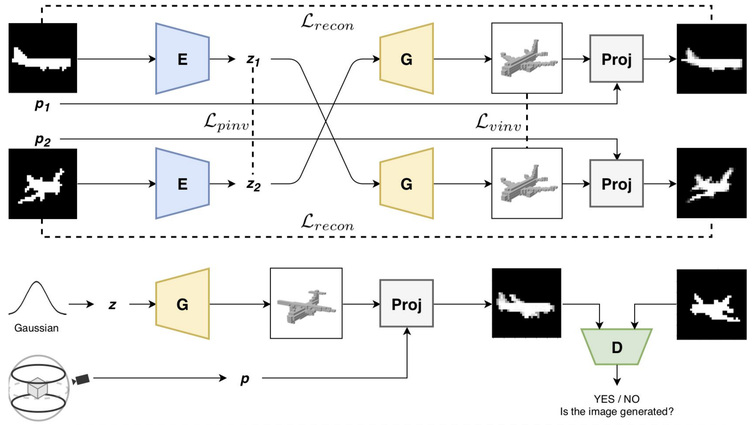
- Most current methods require **camera pose** and **instance IDs**.
- How to best combine *large amount of cheap supervision* with *limited amount of pose annotation*?
- How can we create a unified model for different types of supervision?"

Study through Three Training Paradigms:

- ***Semi-supervised Single Category***
  Limited Instance IDs + Camera Pose + category ID, Unlabeled Images
- ***Semi-supervised Multi-category***
  Limited Instance IDs + Camera Pose, Unlabeled Images
- ***Few-shot Transfer Learning***
  Pretrain: *Semi-supervised Multi-category* setting.
  Train: **Novel category**, *Semi-supervised Single Category setting.*



Architecture



$$\mathcal{L}_{recon} = \|P(G(E(\mathbf{x}_2)), \mathbf{p}_1) - \mathbf{x}_1\|_{1+2} + \|P(G(E(\mathbf{x}_1)), \mathbf{p}_2) - \mathbf{x}_2\|_{1+2}$$

$$\mathcal{L}_{pinv} = \|E(\mathbf{x}_1) - E(\mathbf{x}_2)\|_2 \quad \mathcal{L}_{vinv} = \|G(E(\mathbf{x}_1)) - G(E(\mathbf{x}_2))\|_1$$

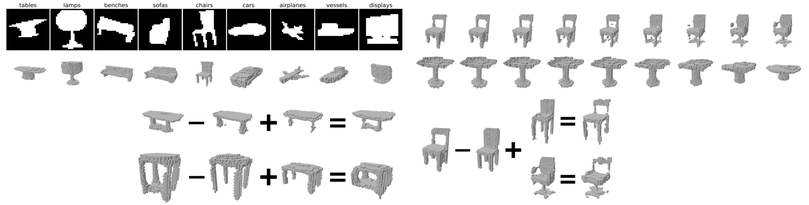
$$\mathcal{L}_D = \mathbb{E}_{\mathbf{z},\mathbf{p}}[\log(1 - D(P(G(\mathbf{z}), \mathbf{p})))] + \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}[\log D(\mathbf{x})]$$

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z},\mathbf{p}}[\log D(P(G(\mathbf{z}), \mathbf{p}))]$$

Single Category Semi-Supervised

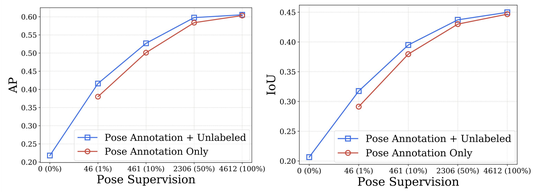
Category	MVC [20] IoU	McRecon [10] AP	McRecon [10] $IoU_{0.4}$	PTN [27] $IoU_{0.5}$	Ours (50% pose annotations) IoU	Ours (50% pose annotations) AP	Ours (50% pose annotations) $IoU_{0.4}$	Ours (50% pose annotations) $IoU_{0.5}$
airplanes	0.55	0.59	0.37	-	**0.57**	**0.75**	**0.56**	0.57
benches	-	0.39	0.30	-	0.36	**0.48**	**0.35**	0.35
cars	0.75	0.82	0.56	-	**0.78**	**0.92**	**0.77**	0.77
chairs	0.42	0.48	0.35	**0.49**	**0.44**	**0.60**	**0.43**	0.42
sofas	-	0.56	0.38	-	0.54	**0.69**	**0.53**	0.52
tables	-	0.46	0.35	-	0.44	**0.63**	**0.43**	0.42

- Using 50% pose with unlabeled data, we can achieve good performance.



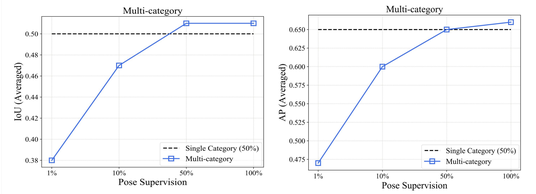
References:
10. Gwak, J., Choy, C.B., Garg, A., Chandraker, M., Savarese, S.: Weakly supervised generative adversarial networks for 3d reconstruction. In: 3DV (2017)
20. Tulsiani, S., Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In: CVPR (2018)
27. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: NIPS (2016)

Are unlabeled images useful?



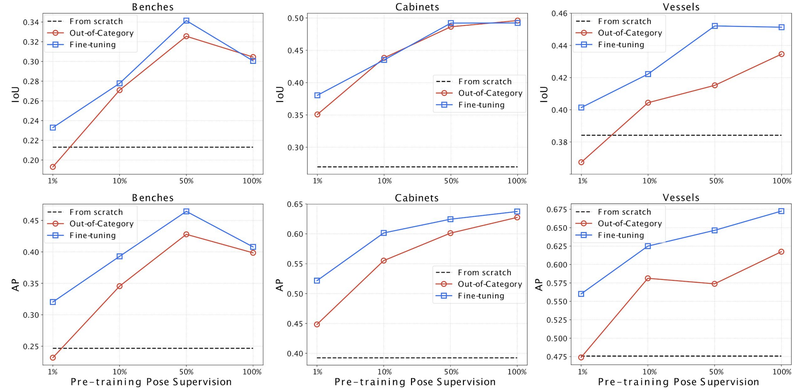
- Adding Unlabeled image improves performance.
- Performance improvement is larger when posed data is more limited.
- Adding even 1% pose annotated data gives large improvement.

Multi-category Semi-Supervised



- Performance in multi-category training is almost the same as of single-category training.
- Category confusion makes this setting harder; mutual information makes this setting easier.
- Category information is hard to leverage to improve the model.

Few-shot Transfer



S - Single category
P - Using pose annotated images
U - Using unlabeled images
M - Multi-categories training
FT - Fine-tuning M

Performances for *chairs* with 1% Pose Annotated Data

	S, P	S, U	S, P+U	M	FT
IoU	0.2913	0.2065	0.3175	0.3104	**0.3250**
AP	0.3800	0.2180	0.4162	0.3859	**0.4247**