View Independent Generative Adversarial Network for Novel View Synthesis

Xiaogang Xu, Ying-Cong Chen, Jiaya Jia
IEEE International Conference on Computer Vision
Presented by: Sean Lee

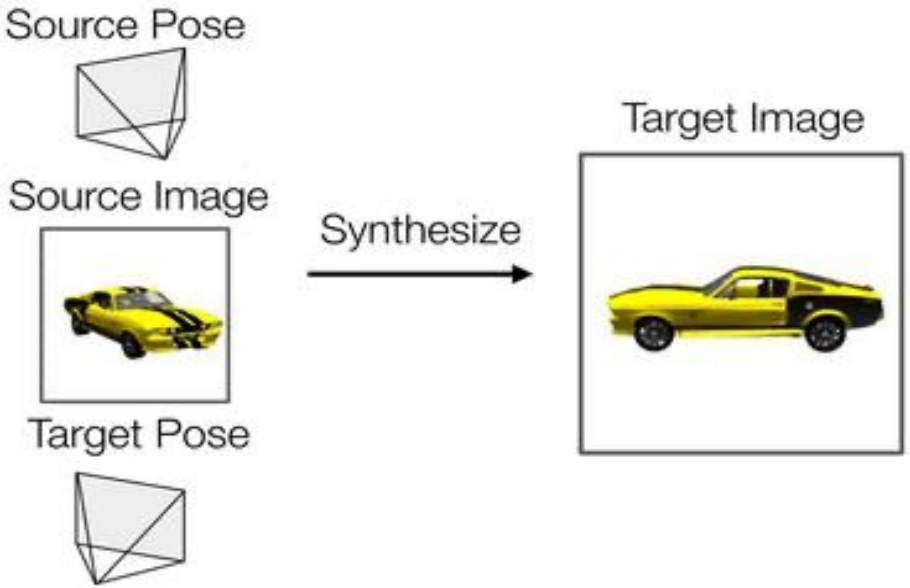
Overview

- Problem
- Related Works
- Motivation
- Proposed Solution
- Evaluation/Results
- Conclusion

Problem

Novel View Synthesis:

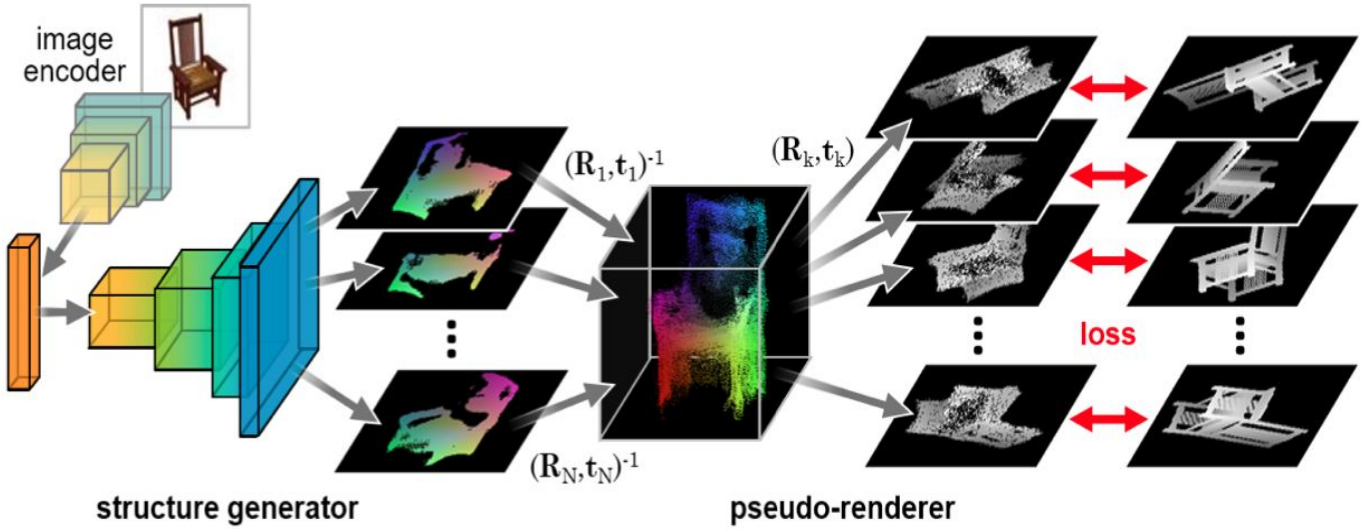
- Given a 2D image synthesize a new 2D image from an arbitrary viewpoint.



Related Works

Geometry based approaches:

Problem: Estimating 3D structures from a 2D image doesn't perform well on complicated structures/textures.

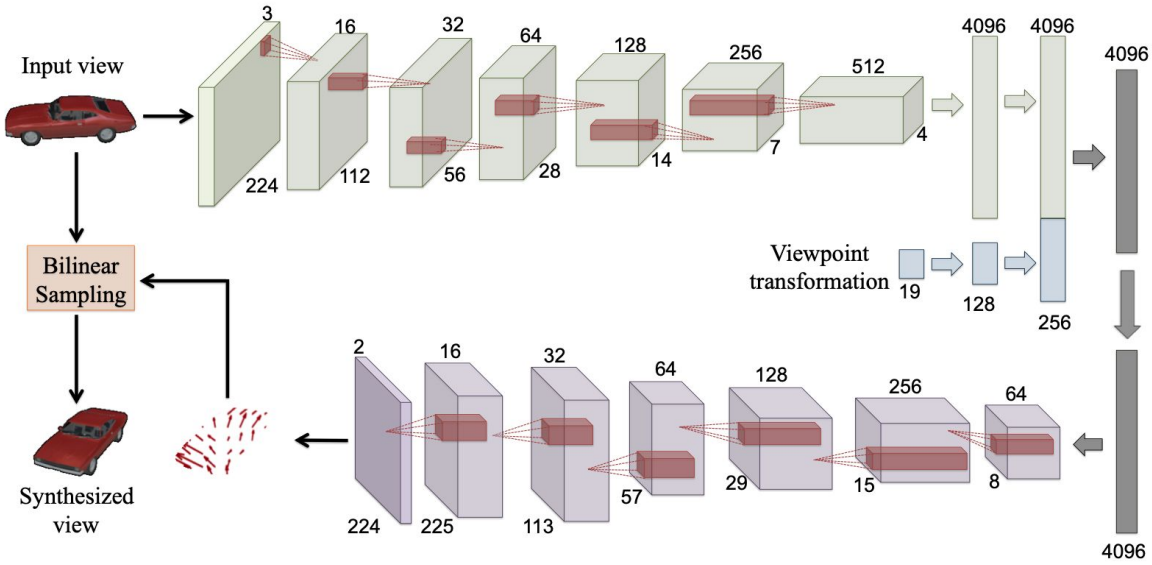


Souce: Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. arXiv e-prints, 2017.

Related Works

Learning based approaches:

Problem: Generating 2D images without considering 3D structures doesn't generalize well.

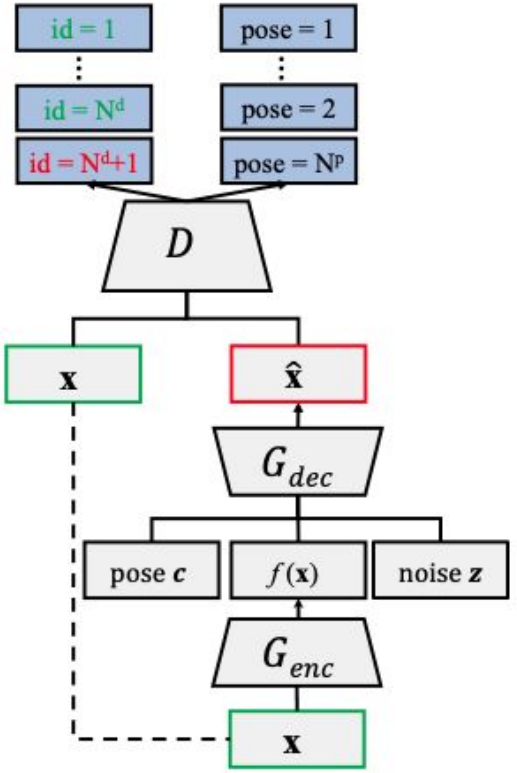


Souce: Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In ECCV, 2016.

Related Works

Generative Adversarial Network:

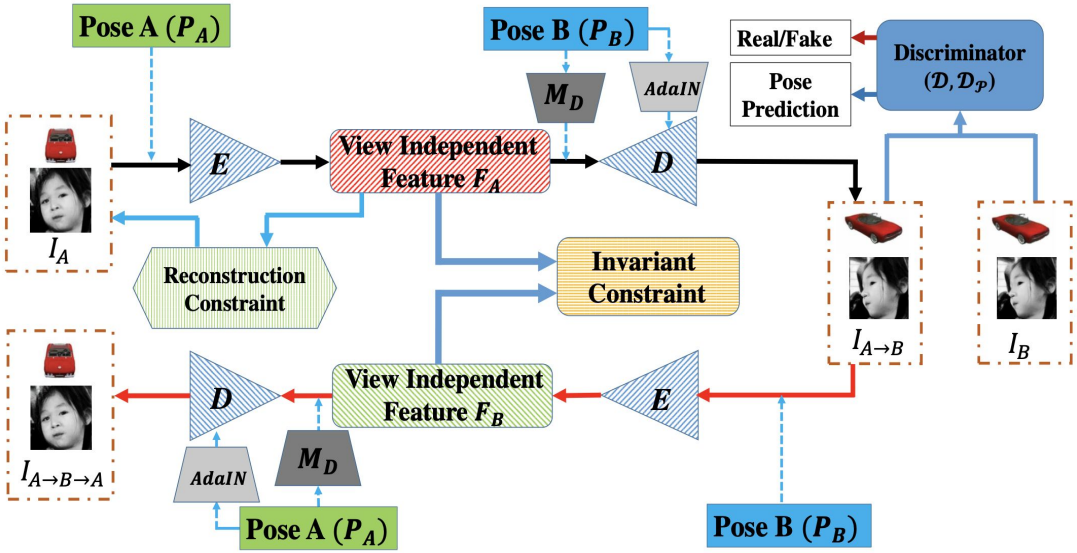
Problem: 3D structures aren't represented. Therefore, the GAN doesn't generalize well.



Souce: Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In CVPR, 2017.

Motivation

Propose an improved method for novel view synthesis.



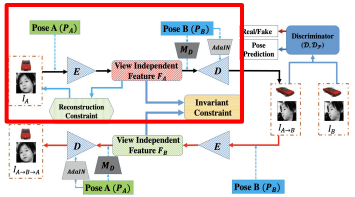
Proposed Solution: Encoder-Decoder based Generative Adversarial Network (VI-GAN).

Proposed Solution

Camera Pose:

- Rotation Matrix: $R \in \mathbb{R}^{3x3}$
- Shifting Vector: $T \in \mathbb{R}^{3x1}$
- R is reshaped (9x1) and concatenate with T (12x1)

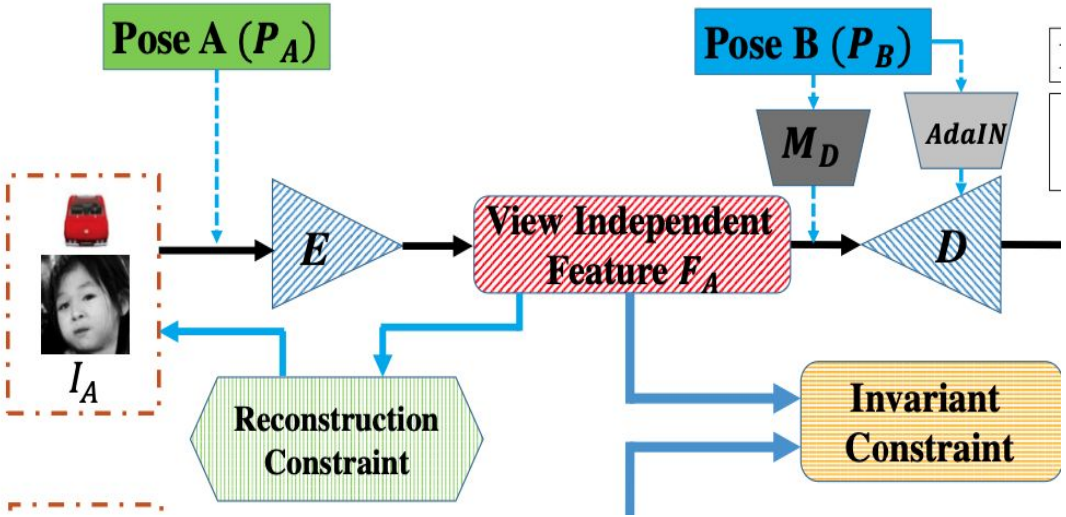
Proposed Solution



Encoder:

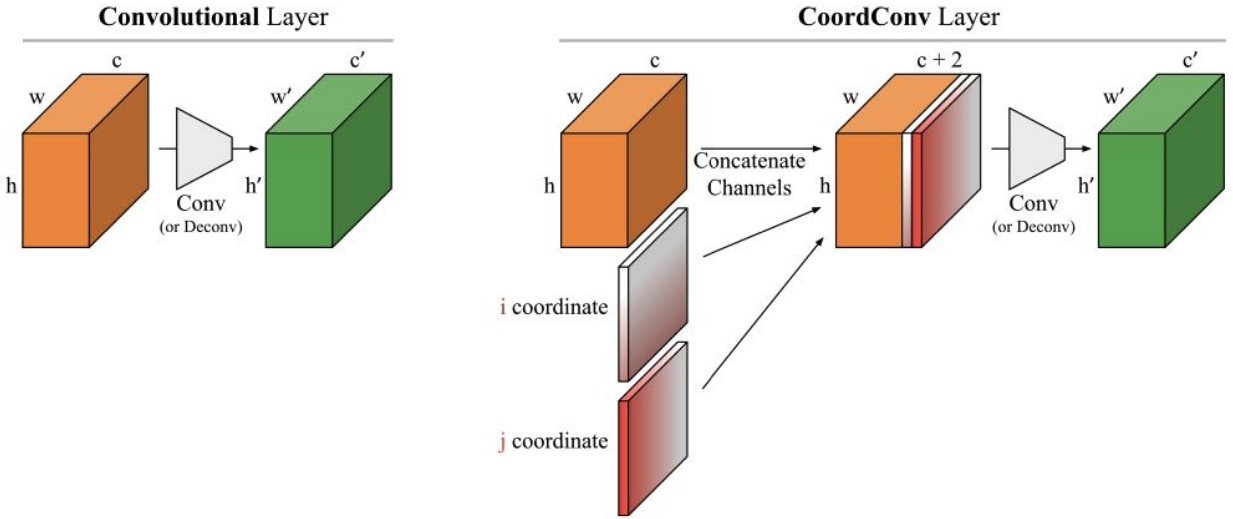
- Input: 2D image (I_A) and camera pose (P_A).
- Output: View-Independent features (F_A).

$$F_A = E(I_A \oplus P_A),$$



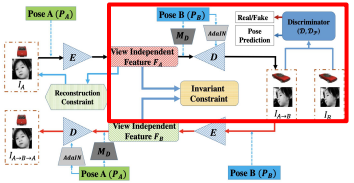
Proposed Solution

CoordCov:



Souce: Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. arXiv e-prints, 2018.

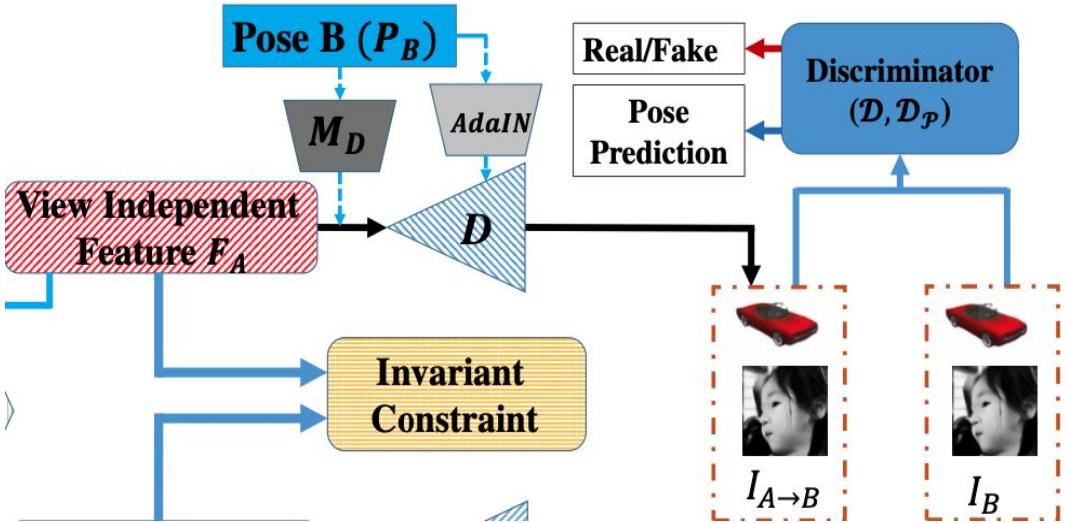
Proposed Solution



Decoder:

- Input:
  View-Independent features (F_A) and target camera pose (P_B).
- Output: Target image ($I_{A->B}$).

$$I_{A\to B} = D(F_A \oplus M_D(P_B)),$$



Proposed Solution

Adaptive Instance Normalization:

- Aligns channel-wise mean and variance of the input (x) and the style (y).
- The mean and variance of the instance normalization layer are inferred by the target pose instead of the feature map.

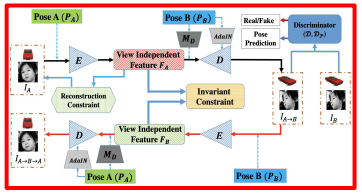
$$\mathbf{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y)$$

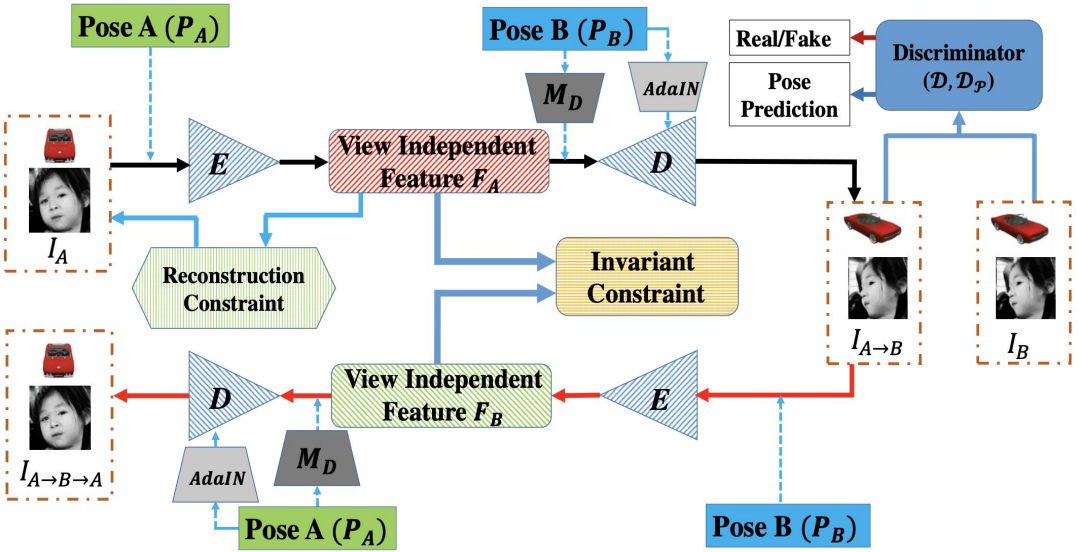
Souce: Xun Huang, Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization.

Proposed Solution

View-Independent Loss:

$$\mathcal{L}_{VI} = \mathbb{E}(\|F_A - F_B\|),$$





Proposed Solution

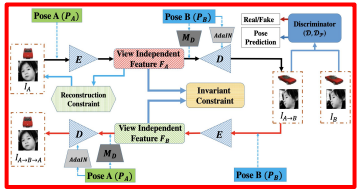


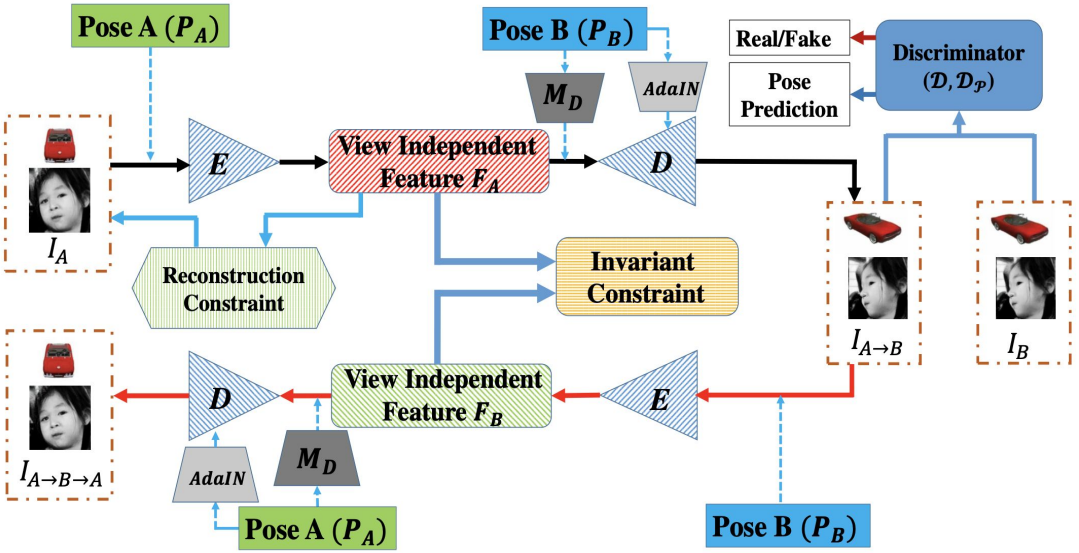
Image-Reconstruction Loss:

- Pixel-level loss:

$$\mathcal{L}_{pixel} = \mathbb{E}(\|I_{A\to B} - I_B\|),$$

- Perceptive loss:

$$\mathcal{L}_{per} = \mathbb{E}((\mathcal{V}(I_{A\to B}) - \mathcal{V}(I_B))^2),$$



Proposed Solution

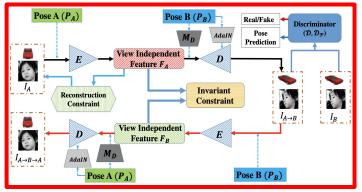


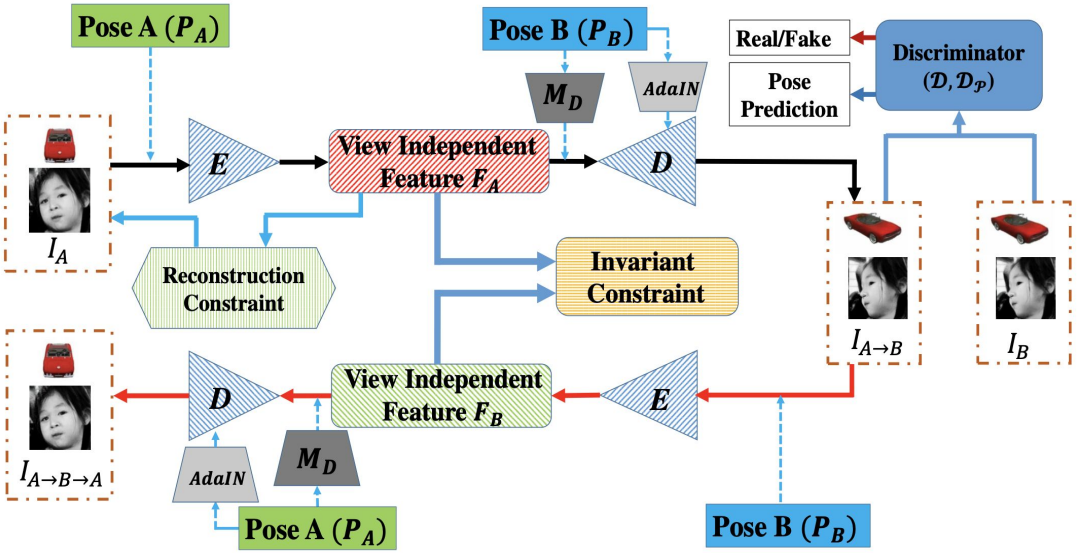
Image-Reconstruction Loss:

- Input reconstruction loss:

$$\mathcal{L}_{rec} = \mathbb{E}(\|I_A - I_{A\to A}\|),$$

- Cycle restriction:

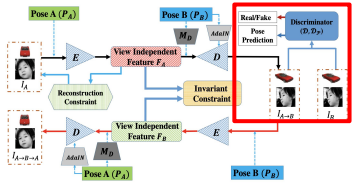
$$\mathcal{L}_{cyc} = \mathbb{E}(\|I_{A\to B\to A} - I_A\|),$$
$$\mathcal{L}_{cyc_{per}} = \mathbb{E}((\mathcal{V}(I_{A\to B\to A}) - \mathcal{V}(I_A))^2).$$

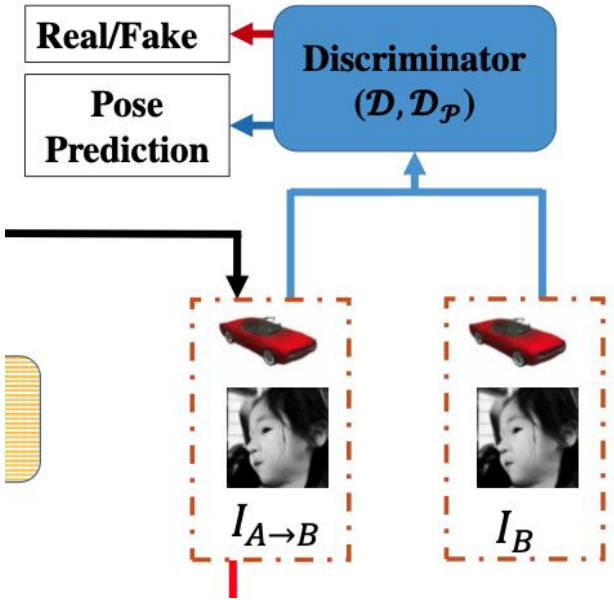


Proposed Solution

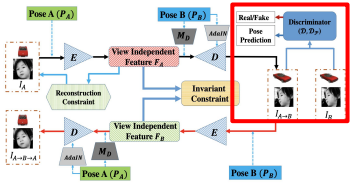
GAN loss (Wasserstein GAN-Gradient Penalty):

$$\mathcal{L}_{\mathcal{GAN}_\mathcal{D}} = \mathbb{E}_{\mathbb{P}_f}[\mathcal{D}(I_{A\to B})] - \mathbb{E}_{\mathbb{P}_r}[\mathcal{D}(I_B)] + \lambda_{gp}\mathcal{L}_{gp},$$
$$\mathcal{L}_{\mathcal{GAN}_\mathcal{G}} = \mathbb{E}_{\mathbb{P}_r}[\mathcal{D}(I_B)] - \mathbb{E}_{\mathbb{P}_f}[\mathcal{D}(I_{A\to B})],$$



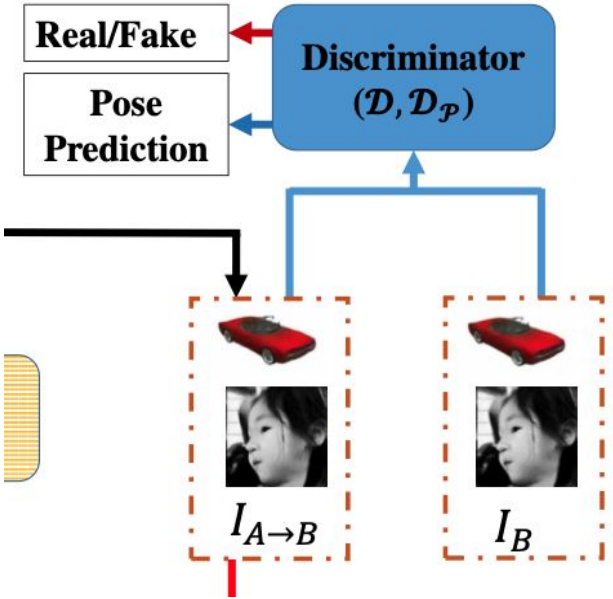


Proposed Solution



Pose Prediction Loss:

$$\mathcal{L}_{\mathcal{GAN}_{\mathcal{D}_\mathcal{P}}} = \mathbb{E}_{\mathbb{P}_r}((\mathcal{D}_\mathcal{P}(I_B) - P_B)^2),$$
$$\mathcal{L}_{\mathcal{GAN}_{\mathcal{G}_\mathcal{P}}} = \mathbb{E}_{\mathbb{P}_f}((\mathcal{D}_\mathcal{P}(I_{A\to B}) - P_B)^2),$$



Evaluation/Results

ShapeNet:



Category	Num	Category	Num	Category	Num	Category	Num	Category	Num
Chair	696	Monitor	127	WallLamp	78	Gun	54	FlagPole	38
Lamp	663	RoundTable	120	SideChair	77	Nightstand	53	TvStand	38
ChestOfDrawers	511	TrashBin	117	VideoGameConsole	75	Mug	51	Fireplace	37
Table	427	DrinkingUtensil	112	MediaStorage	73	AccentChair	50	Rack	37
Couch	413	DeskLamp	110	Painting	73	ChessBoard	49	LightSwitch	36
Computer	244	Clock	101	Desktop	71	Rug	49	Oven	36
Dresser	234	ToyFigure	101	AccentTable	70	WallUnit	46	Airplane	35
TV	233	Plant	98	Camera	70	Mirror	45	DresserWithMirror	35
WallArt	222	Armoire	95	Picture	69	Bowl	44	Calculator	34
Bed	221	QueenBed	94	Refrigerator	68	SodaCan	44	TableClock	34
Cabinet	221	Stool	92	Speaker	68	VideoGameController	44	Toilet	34
FloorLamp	201	EndTable	91	Sideboard	67	WallClock	43	Cup	33
Desk	189	Bottle	88	Barstool	66	Printer	42	Stapler	33
PottedPlant	188	DiningTable	88	Guitar	65	Sword	40	PaperBox	32
FoodItem	180	Bookcase	87	MediaPlayer	62	USBStick	40	SpaceShip	32
Laptop	173	CeilingLamp	86	Ipod	59	Chaise	39	Toy	32
Vase	163	Bench	84	PersonStanding	57	OfficeSideChair	39	ToiletPaper	31
TableLamp	142	Book	84	Piano	56	Poster	39	Knife	30
OfficeChair	137	CoffeeTable	81	Curtain	55	Sink	39	PictureFrame	30
CellPhone	130	Pencil	80	Candle	54	Telephone	39	Recliner	30



Evaluation/Results

Multi-PIE:



	Multi-PIE	PIE
# Subjects	337	68
# Recording Sessions	4	1
High-Resolution Still Images	Yes	No
Calibration Images	Yes	No
# Expressions	6	4
# Cameras	15	13
# Flashes	18	21
Total # Images	750,000+	41,000+
DB Size [GB]	305	40

Evaluation/Results

300W-LP:

- A synthetic expansion of the 300W dataset consisting of 120,000 examples.



Evaluation/Results

Evaluation Metrics:

- Mean pixel-wise error (L1)
- Structural similarity index measure (SSIM)
  - An image quality metric that evaluates the visual impact of lumiance, contrast and structure.
  - Scale of 0.0 to 1.0, where 1.0 means the pictures are identical.
- Frechet Inception Distance (FID)
  - Measures the quality of generated image.
  - The lower FID is, the closer distance the domains of the real and generated images.

Evaluation/Results

Ablation Experiment:

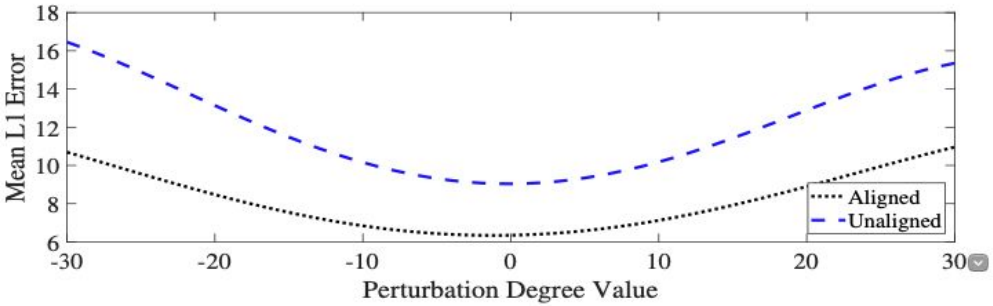
- VI (View-Independent Loss).
- Pose (Pose Prediction Loss).
- Coord (CoordCov).
- AdaIN (Adaptive Instance Normalization).

Setting	L_1	SSIM
VI-GAN(w/o VI)	16.43	0.82
VI-GAN(w/o Pose)	16.81	0.80
VI-GAN(w/o Coord)	14.35	0.85
VI-GAN(w/o AdaIN)	14.02	0.84
VI-GAN	**12.56**	**0.87**

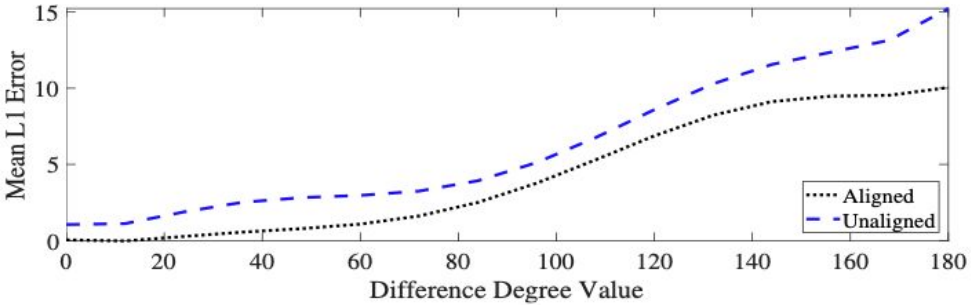
Evaluation/Results

Additional Experiments:

- Camera pose experiment:
  - Tested mean pixel-wise error (L1) for varying pose accuracy and increases in degrees between views.



(a) Pose Sensibility Analysis of Encoder



(b) Pose Sensibility Analysis of Decoder

Evaluation/Results

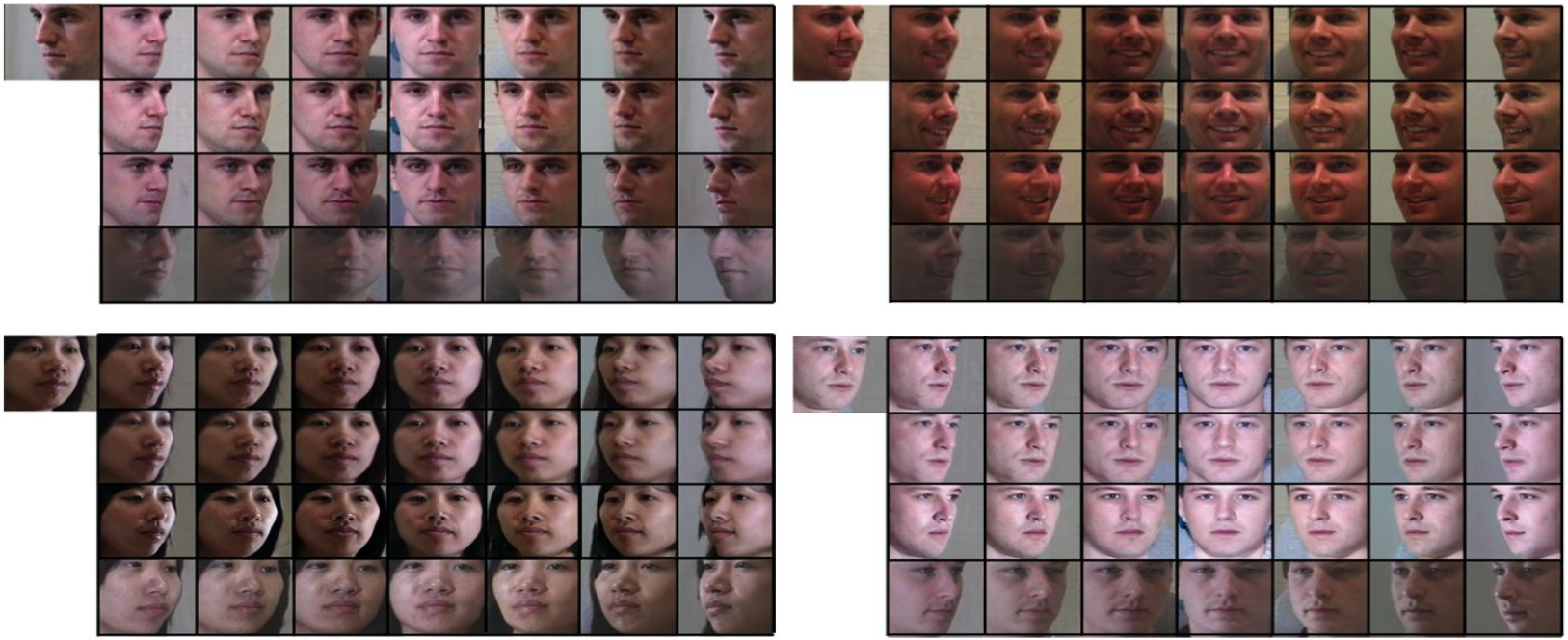
Application:

- Discrete Face Rotation: Face rotation aims to synthesize a human face to another view.

Method	Aligned			Unaligned		
	FID	L_1	SSIM	FID	L_1	SSIM
CR-GAN	8.76	10.17	0.76	13.92	15.45	0.68
DR-GAN	107.5	31.92	0.36	151.1	43.11	0.23
VI-GAN	**6.51**	**5.86**	**0.88**	**9.05**	**9.73**	**0.80**

Evaluation/Results

Discrete Face Rotation (aligned):



Evaluation/Results

Discrete Face Rotation (unaligned):





Evaluation/Results

Application:

- Continuous Face Rotation.

Method	L1	SSIM
PRNet	22.26	0.65
VI-GAN	**15.32**	**0.73**

Evaluation/Results

Continuous Face Rotation:



Evaluation/Results

Application:

- Object Rotation: Synthesize novel views for certain object.

Method	Chair		Sofa		Bench	
	L_1	SSIM	L_1	SSIM	L_1	SSIM
MV3D	24.25	0.76	20.24	0.75	17.52	0.73
AF	18.44	0.82	14.42	0.85	13.26	0.77
VI-GAN	**12.56**	**0.87**	**11.52**	**0.88**	**10.13**	**0.83**

Evaluation/Results

Object Rotation:



(a) Input (b) MV3D (c) AF (d) Ours (e) GT



(a) Input (b) MV3D (c) AF (d) Ours (e) GT

Conclusion

- The paper proposes a technique for utilizing a 3D view synthesis network to generate a 2D target view. The paper demonstrates how the proposed method outperforms existing techniques for:
  - Discrete Face Rotation
  - Continuous Face Rotation
  - Object Rotation

References

- Sun, S.-H., Huh, M., Liao, Y.-H., Zhang, N., & Lim, J. J. (2018). Multi-view to Novel View: Synthesizing Novel Views With Self-learned Confidence. Computer Vision – ECCV 2018 Lecture Notes in Computer Science, 162–178. doi: 10.1007/978-3-030-01219-9_10
- Xiaogang Xu, Ying-Cong Chen, Jiaya Jia. View Independent Generative Adversarial Network for Novel View Synthesis.
- Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. arXiv e-prints, 2017.
- Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In ECCV, 2016.
- Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In CVPR, 2017.
- Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. arXiv e-prints, 2018.
- Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository
- Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, Simon Baker. Multi-PIE
- Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, Stan Z. Li. Face Alignment Across Large Poses: A 3D Solution
- Xun Huang, Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization.

Questions?