Visual Learning with Weak Supervision

Safa Cicek

Committee:
Prof. Stefano Soatto, Chair
Prof. Lieven Vandenberghe
Prof. Paulo Tabuada
Prof. Guy Van den Broeck

PHD Defense

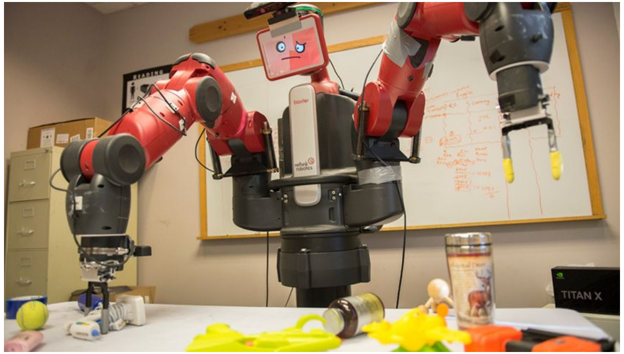
January 2021

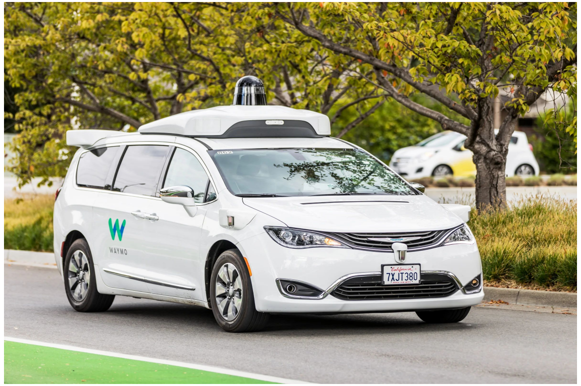
UCLA ELECTRICAL AND COMPUTER ENGINEERING

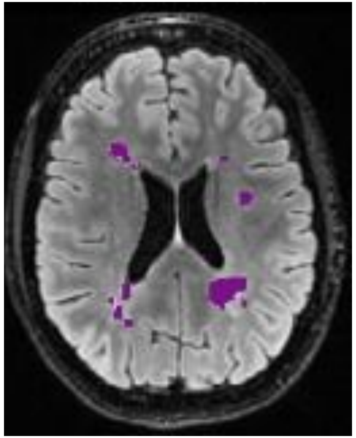
Table of Contents

[1] Cicek, Safa, Alhussein Fawzi, and Stefano Soatto. Saas: Speed as a supervisor for semi-supervised learning. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[2] Cicek, Safa, and Stefano Soatto. Input and Weight Space Smoothing for Semi-supervised Learning. Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops. 2019.
[3] Cicek, Safa, and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019.
[4] Cicek, Safa, Zhaowen Wang, Hailin Jin, Stefano Soatto, Generative Feature Disentangling for Unsupervised Domain Adaptation. *Proceedings of the European Conference on Computer Vision (ECCV)* Workshops. (2020).
[5] Cicek, Safa, Ning Xu, Zhaowen Wang, Hailin Jin, Stefano Soatto, Spatial Class Distribution Shift in Unsupervised Domain Adaptation. Asian Conference on Computer Vision (ACCV). 2020.
[6] Wong Alex, Safa Cicek, Stefano Soatto, Learning Topology from Synthetic Data for Unsupervised Depth Completion, IEEE Robotics and Automation Letters (RAL). 2021.
[7] Wong Alex, Safa Cicek, Stefano Soatto, Targeted Adversarial Perturbations for Monocular Depth Prediction. Conference on Neural Information Processing Systems (NeurIPS). 2020.

Visual Perception







[1] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.

Manual annotation is expensive.

Image Classification

[1]

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

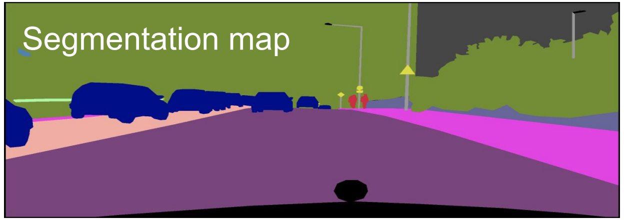
Manual annotation is expensive.

Image Classification

[1]

Semantic Segmentation

[2]



[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
[2] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

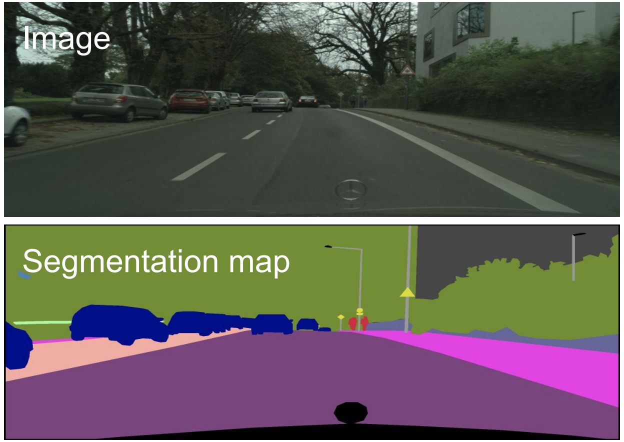
Manual annotation is expensive.

Image Classification

[1]

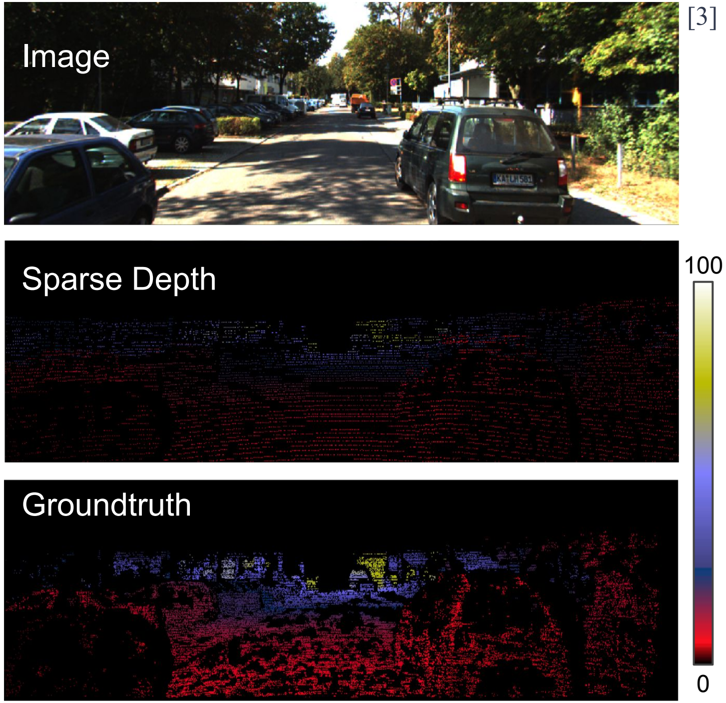
Semantic Segmentation

[2]



Sparse to Dense Depth Completion

[3]

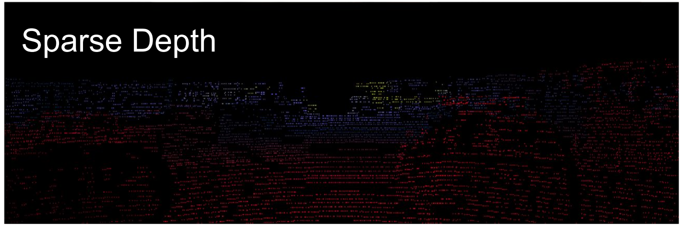


[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
[2] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[3] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant cnns. 3DV 2017.

Unlabeled Real Data

[1]

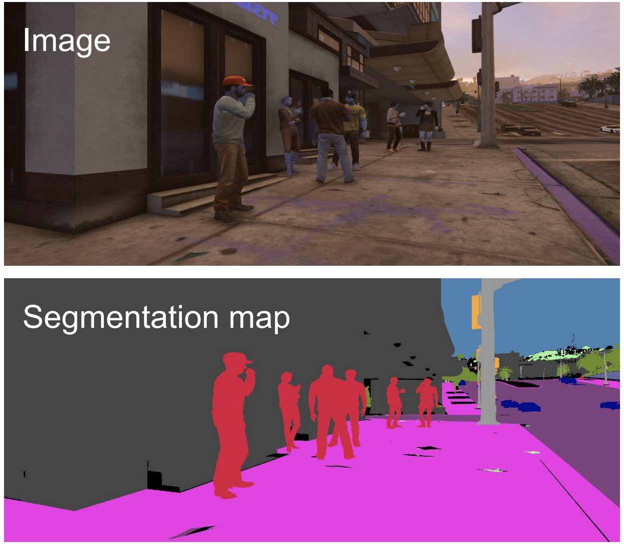
[2]



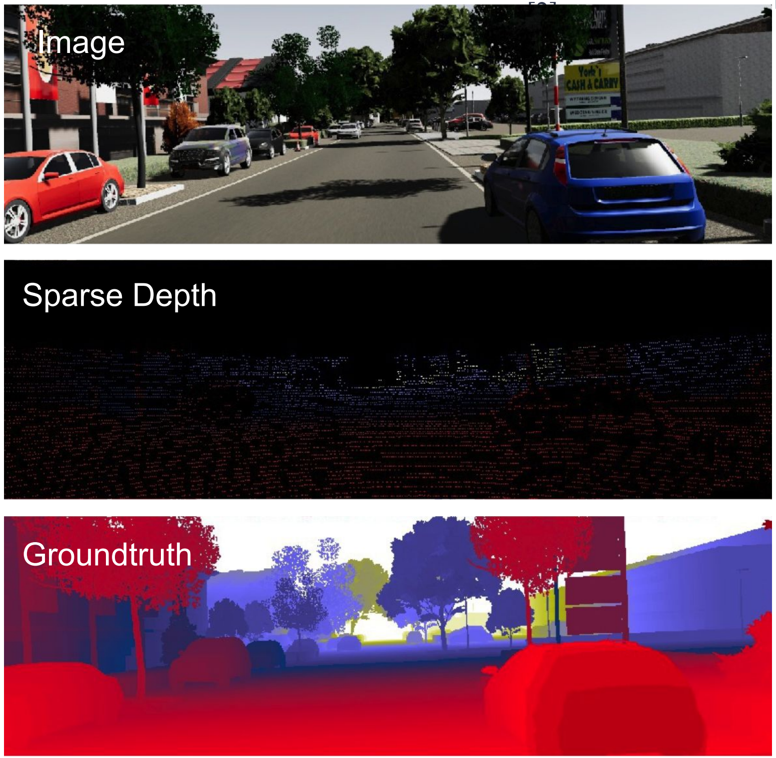
[1] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[2] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant cnns. 3DV 2017.

Unlabeled Real Data + Labeled Virtual Data

[1]



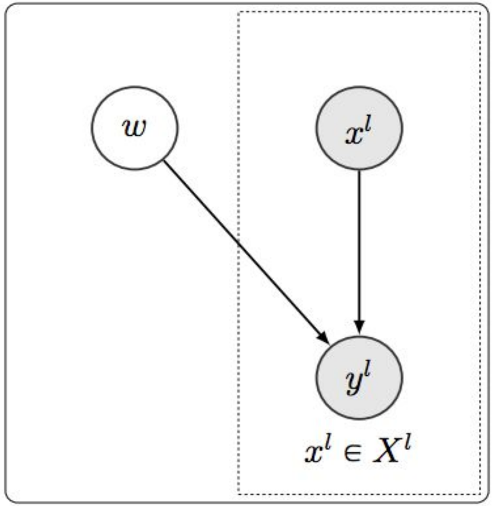
[2]



[1] Richter, Stephan R., et al. "Playing for data: Ground truth from computer games." European conference on computer vision. Springer, Cham, 2016.
[2] Y. Cabon, N. Murray, M. Humenberger. Virtual KITTI 2. Preprint 2020.

Dependency of Unlabeled Data Labels and Model Parameters

Discriminative supervised

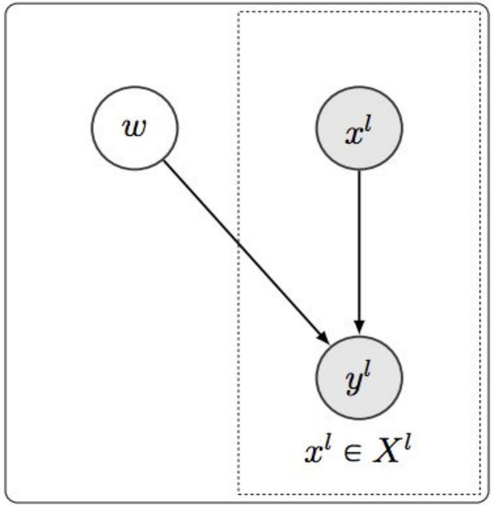


- Shaded variables are fully observed.

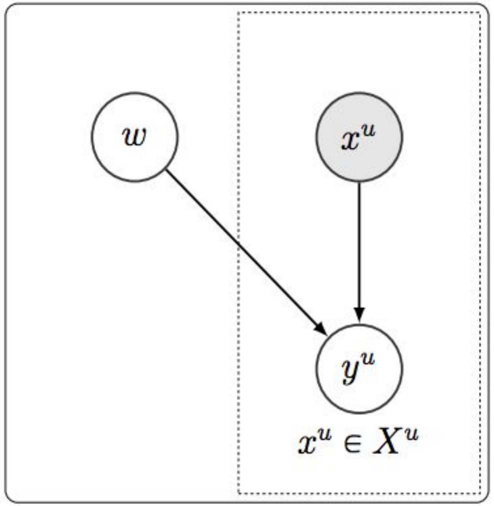
[1] Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)." IEEE Transactions on Neural Networks 20.3 (2009): 542-542.
[2] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Dependency of Unlabeled Data Labels and Model Parameters

Discriminative supervised



Discriminative unsupervised

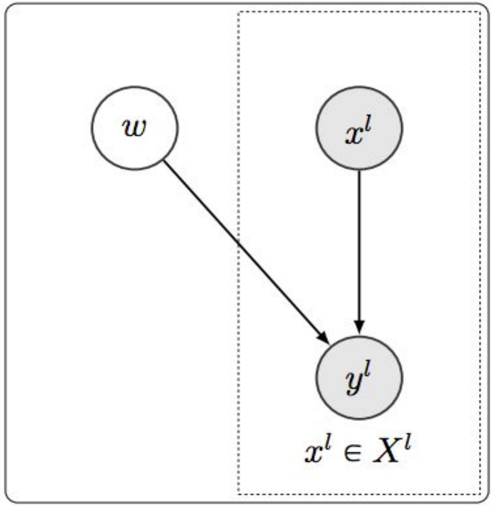


- Shaded variables are fully observed.

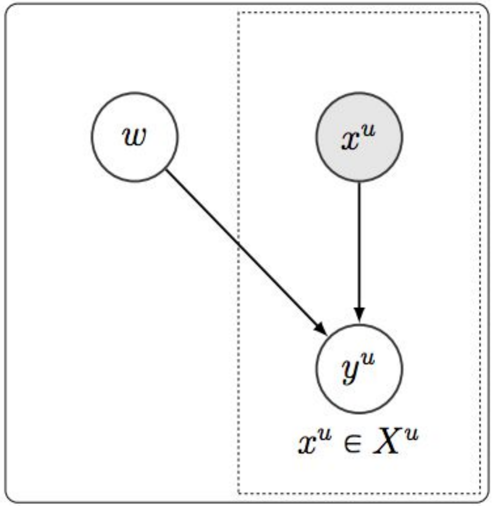
[1] Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)." IEEE Transactions on Neural Networks 20.3 (2009): 542-542.
[2] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Dependency of Unlabeled Data Labels and Model Parameters

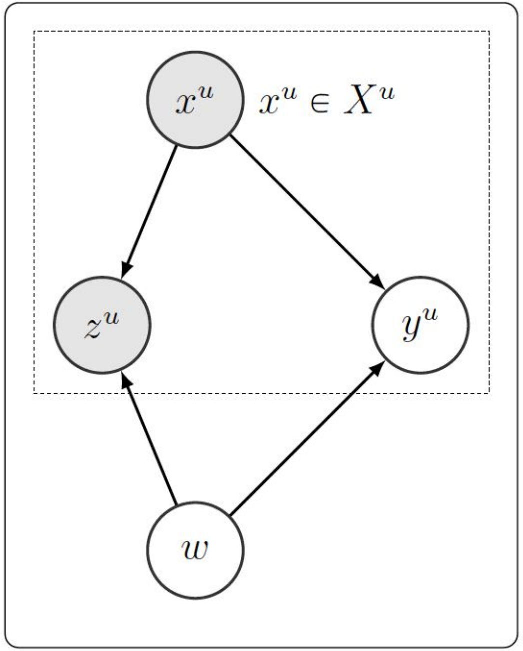
Discriminative supervised



Discriminative unsupervised



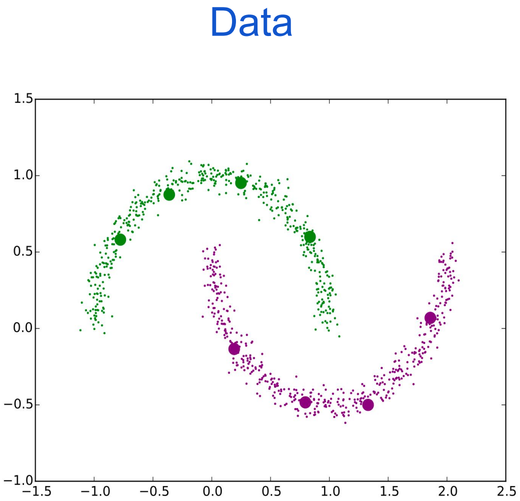
Discriminative Regularized



- Shaded variables are fully observed.

[1] Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)." IEEE Transactions on Neural Networks 20.3 (2009): 542-542.
[2] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Max-margin (Cluster, Low-density) Assumption

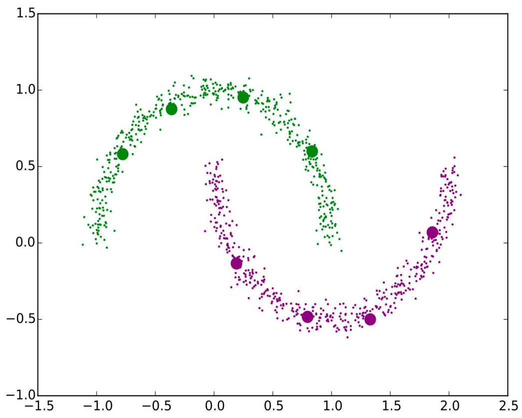


- Large circles (4+4) are labeled samples.

- Small dots are unlabeled samples.

Max-margin (Cluster, Low-density) Assumption

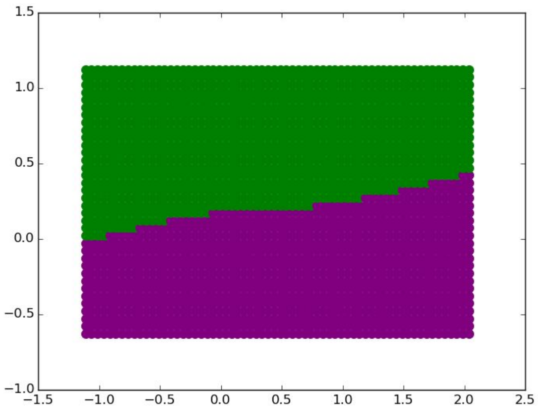
Data



- Large circles (4+4) are labeled samples.

- Small dots are unlabeled samples.

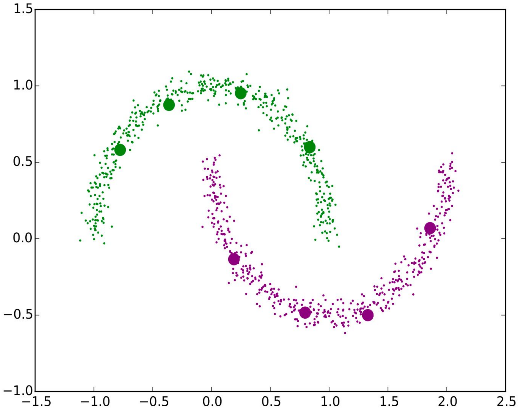
Learned Decision Boundaries



- Without regularization, only using labeled samples.

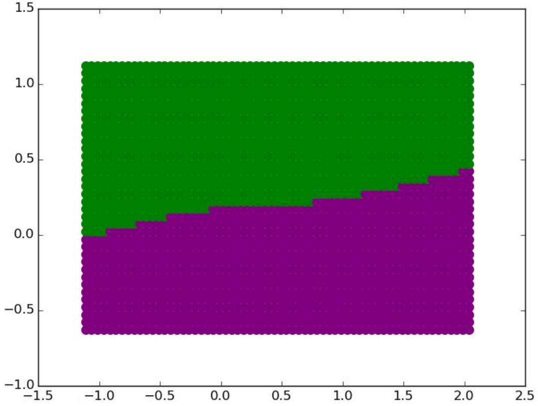
Max-margin (Cluster, Low-density) Assumption

Data

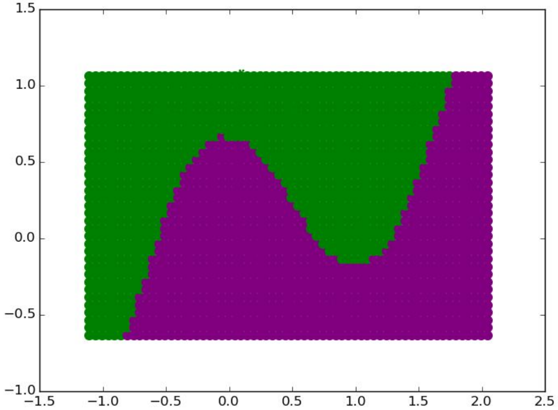


- Large circles (4+4) are labeled samples.
- Small dots are unlabeled samples.

Learned Decision Boundaries



- Without regularization, only using labeled samples.



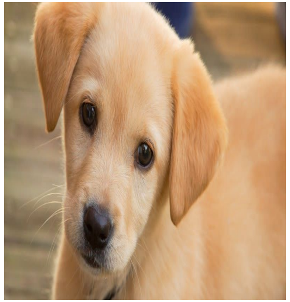
- With regularization (e.g. VAT [1]), also using unlabeled samples.

[1] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2017). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976.

SaaS: Speed as a Supervisor for Semi-supervised Learning

[1] Cicek, Safa, Alhussein Fawzi, and Stefano Soatto. Saas: Speed as a supervisor for semi-supervised learning. Proceedings of the European Conference on Computer Vision (ECCV). 2018.

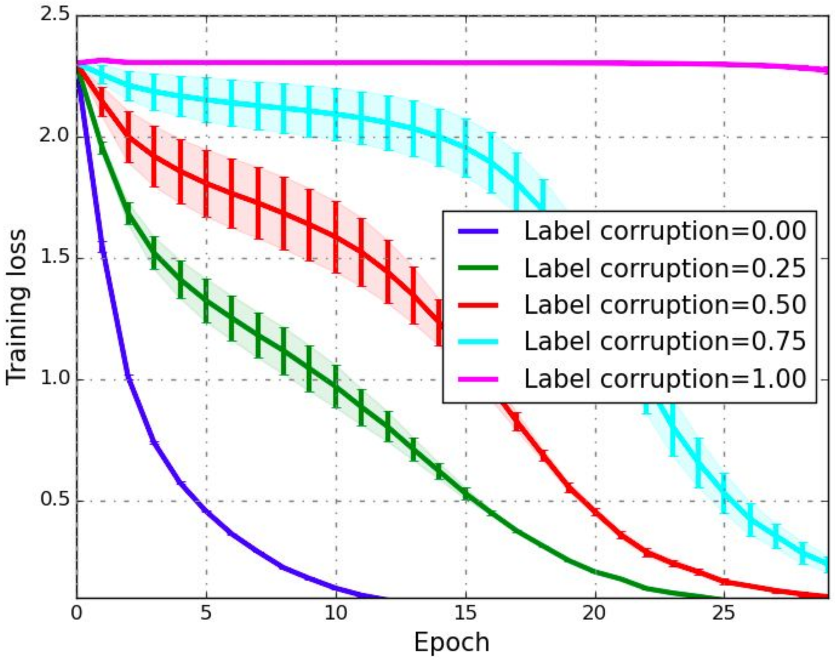
Semi-supervised Learning



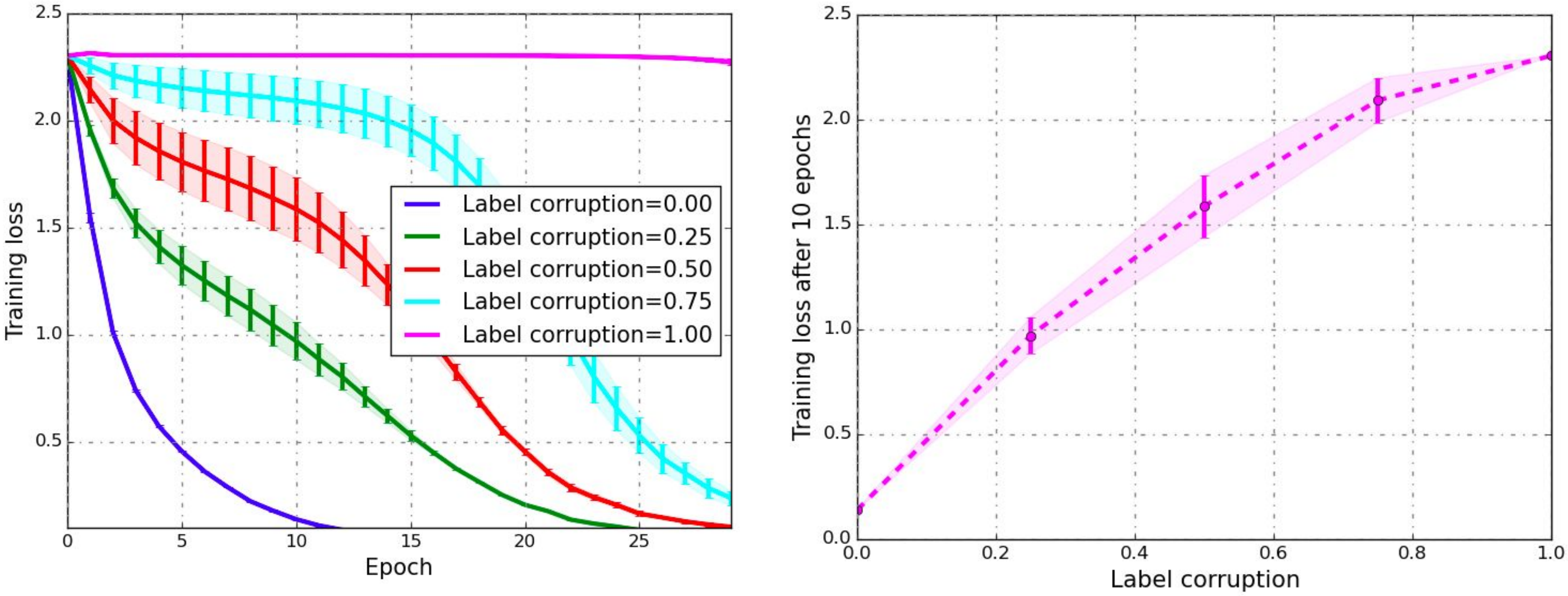
$$\underbrace{\overbrace{\{x_i^l\}_{i=1}^{N^l} \quad \{y_i^l\}_{i=1}^{N^l}}^{\text{Supervised}} \quad \{x_i^u\}_{i=1}^{N^u}}_{\text{Semi-supervised}}$$

$$f_w(x) \in \mathbb{R}^K$$

SaaS: Speed as a Supervisor for Semi-supervised Learning



SaaS: Speed as a Supervisor for Semi-supervised Learning



SaaS

- Inner loop to measure ease of training for the current pseudo-labels.

- Outer loop to update the pseudo-labels.

$P^u \sim \mathcal{N}(0, I)$

Select learning rates η for the weights η_w and label posteriors η_{P^u}

Phase I: Estimate P^u

while P^u has not stabilized **do**

$P^u = \Pi(P^u)$ (project posterior onto the probability simplex)

$w_1 \sim \mathcal{N}(0, I)$

$\Delta P^u = 0$

// Run SGD for T steps (on the weights) to estimate loss decrease

for $t = 1 : T$ **do**

$$w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \left(\ell(B_t^u, P^u; w_{t-1}) + \beta q(B_t^u; w_{t-1})\right)$$

$$w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \ell(B_t^l, P^l; w_{t-\frac{1}{2}})$$

$$\Delta P^u = \Delta P^u + \nabla_{P^u} \ell(B_t^u, P^u; w_t)$$

// Update the posterior distribution

$P^u = P^u - \eta_{P^u} \Delta P^u$

Phase II: Estimate the weights.

$\hat{y}_i^u = \arg\max_i P_i^u \ \forall i = 1, \ldots, N^u$

$w_1 \sim \mathcal{N}(0, I)$

while w has not stabilized **do**

$$w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \frac{1}{|B_t^u|} \sum_{i=1}^{|B_t^u|} \ell(x_i^u, \hat{y}_i^u; w_{t-1})$$

$$w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \frac{1}{|B_t^l|} \sum_{i=1}^{|B_t^l|} \ell(x_i^l, y_i^l; w_{t-\frac{1}{2}})$$

SaaS

- Inner loop to measure ease of training for the current pseudo-labels.

$P^u \sim \mathcal{N}(0, I)$
Select learning rates η for the weights η_w and label posteriors η_{P^u}
Phase I: Estimate P^u
while P^u has not stabilized **do**
    $P^u = \Pi(P^u)$ (project posterior onto the probability simplex)
    $w_1 \sim \mathcal{N}(0, I)$
    $\Delta P^u = 0$
    // Run SGD for T steps (on the weights) to estimate loss decrease
    for $t = 1 : T$ **do**
        $w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \left(\ell(B_t^u, P^u; w_{t-1}) + \beta q(B_t^u; w_{t-1})\right)$
        $w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \ell(B_t^l, P^l; w_{t-\frac{1}{2}})$
        $\Delta P^u = \Delta P^u + \nabla_{P^u} \ell(B_t^u, P^u; w_t)$
    // Update the posterior distribution
    $P^u = P^u - \eta_{P^u} \Delta P^u$
Phase II: Estimate the weights.
$\hat{y}_i^u = \arg\max_i P_i^u \; \forall i = 1, \ldots, N^u$
$w_1 \sim \mathcal{N}(0, I)$
while w has not stabilized **do**
    $w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \frac{1}{|B_t^u|} \sum_{i=1}^{|B_t^u|} \ell(x_i^u, \hat{y}_i^u; w_{t-1})$
    $w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \frac{1}{|B_t^l|} \sum_{i=1}^{|B_t^l|} \ell(x_i^l, y_i^l; w_{t-\frac{1}{2}})$

Objective Function

$$\mathcal{L}_T(P^u) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|B_t^u|}\sum_{i=1}^{|B_t^u|} - \underbrace{\langle \log f_{w_t}(x_i^u), P_i^u \rangle}_{\ell(x_i^u, P_i^u; w_t)}$$

$$P^u \in \mathbb{R}^{N^u \times K}$$

$$P_i^u[k] = P(y_i = k | x_i), \;\; k = 1, \ldots, K$$

$$P^u = \arg\min_{P^u} \quad \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \underbrace{\ell(B_t^u, P^u; w_{t-1})}_{\frac{1}{|B_t^u|}\sum_{i=1}^{|B_t^u|}\ell(g_i(x_i^u), P_i^u; w_{t-1})}$$

- Cumulative loss: area under the loss curve up to a small number of epochs.

Degenerate Solutions to Cumulative Loss

- Supervision quality correlates with learning speed in *expectation* not in every *realization*.

Degenerate Solutions to Cumulative Loss

- Supervision quality correlates with learning speed in *expectation* not in every *realization*.
    - Posterior of label estimates should live in probability simplex.
    - Entropy minimization [1,2]
    - Cumulative loss should be small for augmented unlabeled data.

$$P^u \in \mathcal{S}$$

$$H_Q(w) = \sum_{i=1}^{N^u} - \underbrace{\langle f_w(x_i^u), \log f_w(x_i^u) \rangle}_{q(x_i^u; w)}$$

[1] Grandvalet, Yves, and Yoshua Bengio. "Semi-supervised learning by entropy minimization." Advances in neural information processing systems. 2005.
[2] Krause, Andreas, Pietro Perona, and Ryan G. Gomes. "Discriminative clustering by regularized information maximization." Advances in neural information processing systems. 2010.

Degenerate Solutions to Cumulative Loss

- Supervision quality correlates with learning speed in *expectation* not in every *realization*.
    - Posterior of label estimates should live in probability simplex.
    - Entropy minimization [1,2]
    - Cumulative loss should be small for augmented unlabeled data.
    - A strong network can fit to completely random labels [3].
        - So, we measure the speed after a few epochs of training.

$$P^u \in \mathcal{S}$$

$$H_Q(w) = \sum_{i=1}^{N^u} - \underbrace{\langle f_w(x_i^u), \log f_w(x_i^u) \rangle}_{q(x_i^u; w)}$$

$$\underbrace{\min_{w, P^u} \sum_{i=1}^{N} \ell(x_i, P_i^u; w)}$$

This is not equivalent to our optimization.

[1] Grandvalet, Yves, and Yoshua Bengio. "Semi-supervised learning by entropy minimization." Advances in neural information processing systems. 2005.
[2] Krause, Andreas, Pietro Perona, and Ryan G. Gomes. "Discriminative clustering by regularized information maximization." Advances in neural information processing systems. 2010.
[3] Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." arXiv preprint arXiv:1611.03530 (2016).

SaaS

- Outer loop to update the pseudo-labels.

$P^u \sim \mathcal{N}(0, I)$

Select learning rates η for the weights η_w and label posteriors η_{P^u}

Phase I: Estimate P^u

while P^u has not stabilized **do**

$P^u = \Pi(P^u)$ (project posterior onto the probability simplex)

$w_1 \sim \mathcal{N}(0, I)$

$\Delta P^u = 0$

// Run SGD for T steps (on the weights) to estimate loss decrease

for $t = 1 : T$ **do**

$w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \left(\ell(B_t^u, P^u; w_{t-1}) + \beta q(B_t^u; w_{t-1})\right)$

$w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \ell(B_t^l, P^l; w_{t-\frac{1}{2}})$

$\Delta P^u = \Delta P^u + \nabla_{P^u} \ell(B_t^u, P^u; w_t)$

// Update the posterior distribution

$P^u = P^u - \eta_{P^u} \Delta P^u$

Phase II: Estimate the weights.

$\hat{y}_i^u = \arg\max_i P_i^u \ \forall i = 1, \ldots, N^u$

$w_1 \sim \mathcal{N}(0, I)$

while w has not stabilized **do**

$w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \frac{1}{|B_t^u|} \sum_{i=1}^{|B_t^u|} \ell(x_i^u, \hat{y}_i^u; w_{t-1})$

$w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \frac{1}{|B_t^l|} \sum_{i=1}^{|B_t^l|} \ell(x_i^l, y_i^l; w_{t-\frac{1}{2}})$

SaaS

$P^u \sim \mathcal{N}(0, I)$
Select learning rates η for the weights η_w and label posteriors η_{P^u}
Phase I: Estimate P^u
while P^u has not stabilized **do**
  $P^u = \Pi(P^u)$ (project posterior onto the probability simplex)
  $w_1 \sim \mathcal{N}(0, I)$
  $\Delta P^u = 0$
  // Run SGD for T steps (on the weights) to estimate loss decrease
  for $t = 1 : T$ **do**
    $w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \left(\ell(B_t^u, P^u; w_{t-1}) + \beta q(B_t^u; w_{t-1})\right)$
    $w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \ell(B_t^l, P^l; w_{t-\frac{1}{2}})$
    $\Delta P^u = \Delta P^u + \nabla_{P^u} \ell(B_t^u, P^u; w_t)$
  // Update the posterior distribution
  $P^u = P^u - \eta_{P^u} \Delta P^u$

Phase II: Estimate the weights.
$\hat{y}_i^u = \arg\max_i P_i^u \ \forall i = 1, \ldots, N^u$
$w_1 \sim \mathcal{N}(0, I)$
while w has not stabilized **do**
  $w_{t-\frac{1}{2}} = w_{t-1} - \eta_w \nabla_{w_{t-1}} \frac{1}{|B_t^u|} \sum_{i=1}^{|B_t^u|} \ell(x_i^u, \hat{y}_i^u; w_{t-1})$
  $w_t = w_{t-\frac{1}{2}} - \eta_w \nabla_{w_{t-\frac{1}{2}}} \frac{1}{|B_t^l|} \sum_{i=1}^{|B_t^l|} \ell(x_i^l, y_i^l; w_{t-\frac{1}{2}})$

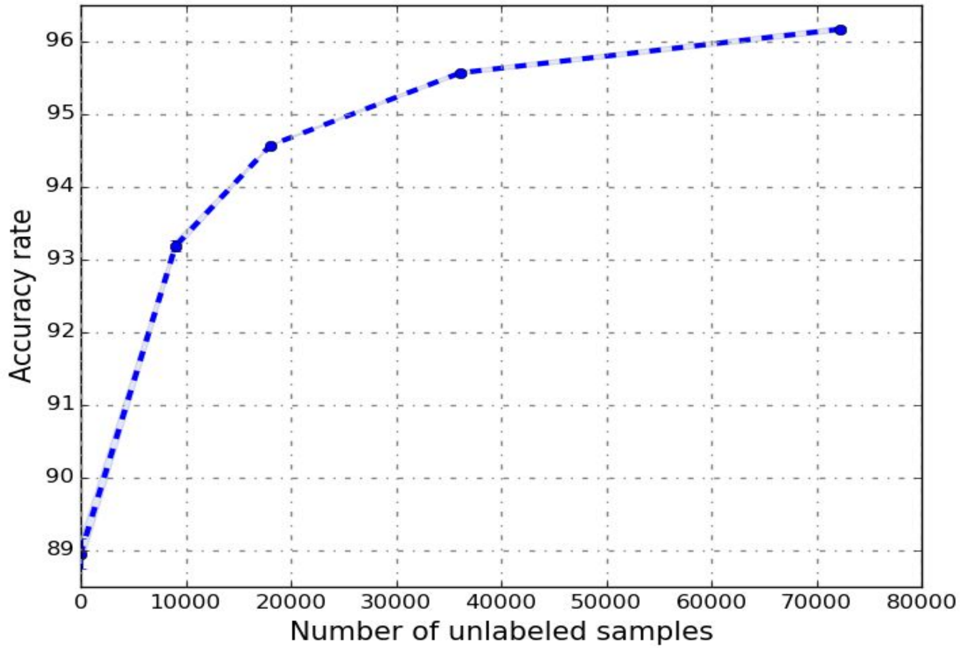
- Learn the model weights from the final pseudo-labels.

Empirical Evaluations

	CIFAR10-4K	SVHN-1K
Error rate by supervised baseline on test data	17.64 ± 0.58	11.04 ± 0.50
Error rate by SaaS on unlabeled data	12.81 ± 0.08	6.22 ± 0.02
Error rate by SaaS on test data	10.94 ± 0.07	3.82 ± 0.09

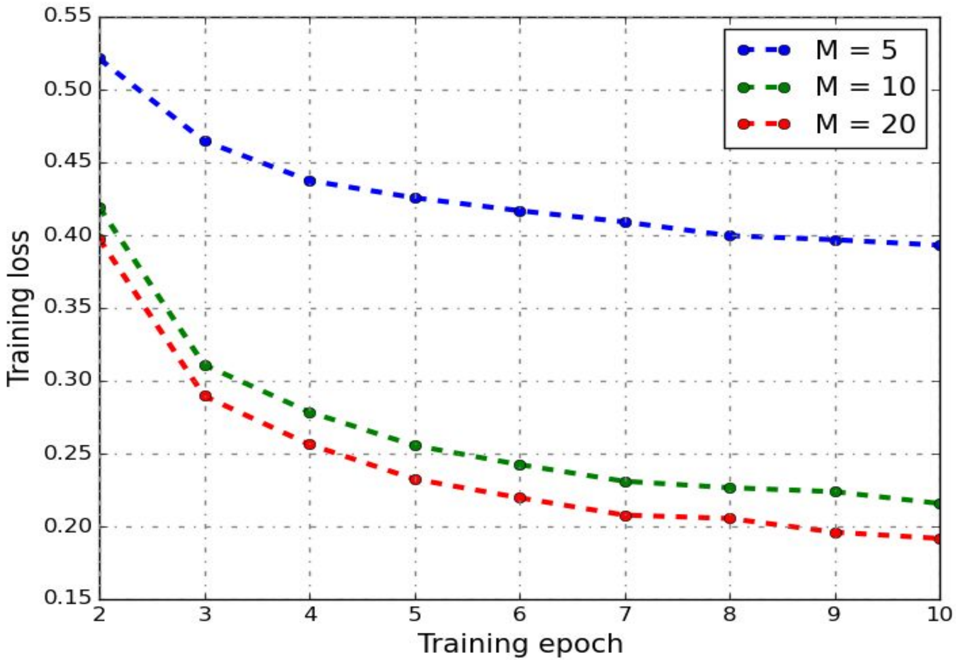
- Comparison to the baseline.

Empirical Evaluations



- The more unlabeled data the better generalization.

Empirical Evaluations



- M is the number of pseudo-label updates.
- SaaS finds labels training on which is faster.

Empirical Evaluations

	Mean Teacher [1]	VAT [2]	SaaS
SVHN-1K	3.95	3.86	**3.82** ± 0.09
CIFAR-4K	12.31	**10.55**	10.94 ± 0.07

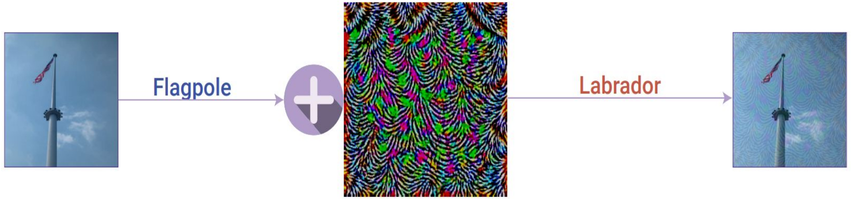
- Comparison to state of the art.

[1] Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.
[2] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2017). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976.

Input and Weight Space Smoothing for Semi-supervised Learning

[1] Cicek, Safa, and Stefano Soatto. Input and Weight Space Smoothing for Semi-supervised Learning. Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops. 2019.

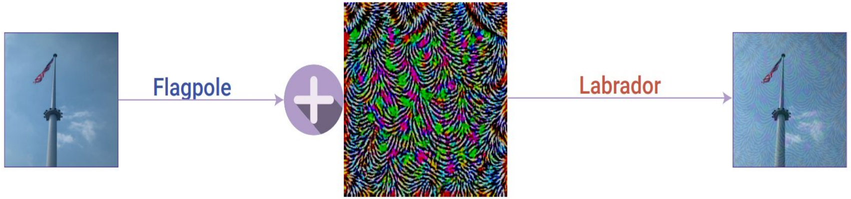
Motivation for Input and Weight Space Smoothing



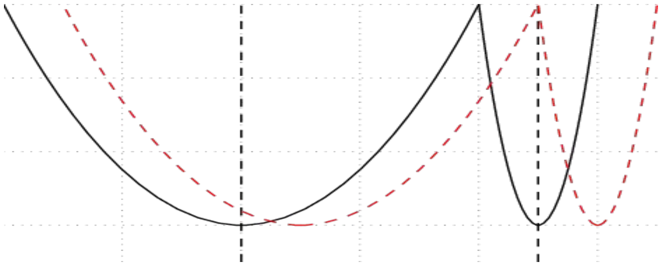
Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

- Small adversarial perturbations are nuisances for the tasks that we are interested in.

Motivation for Input and Weight Space Smoothing



Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.



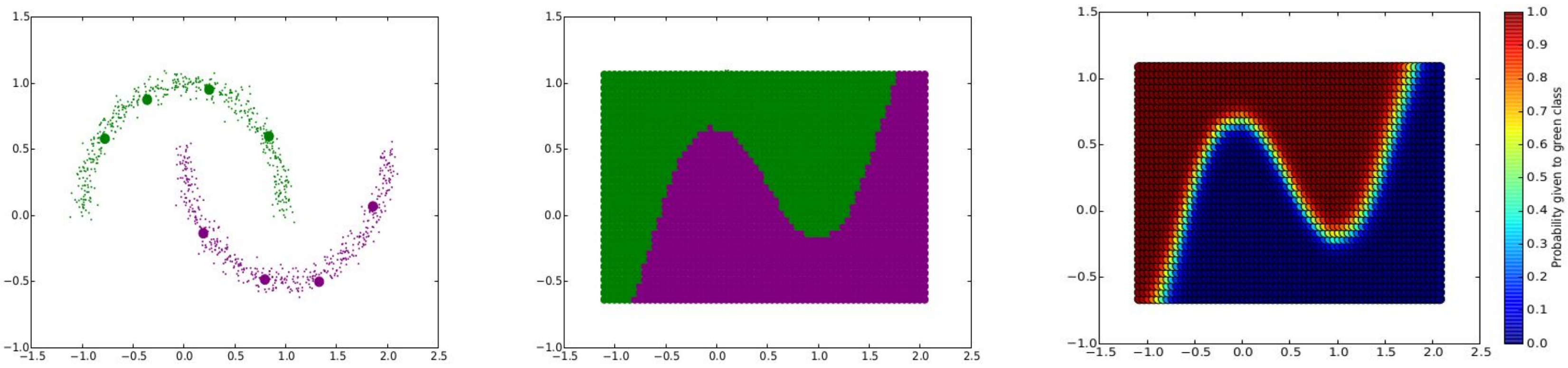
Keskar, N. S., et. al.. (2016). On large-batch training for deep learning: Generalization gap and sharp minima.

- Small adversarial perturbations are nuisances for the tasks that we are interested in.

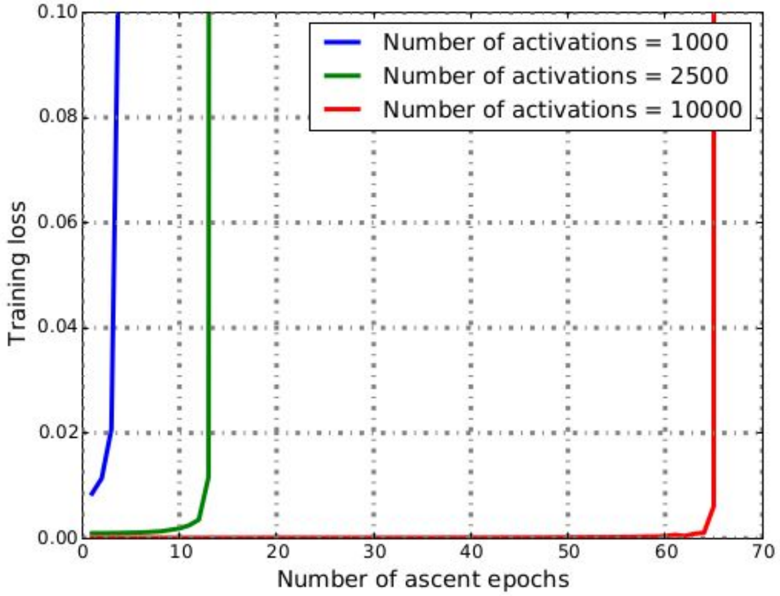
- Converging to a flat-minimum improves generalization [1, 2].

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. "Flat minima." *Neural Computation* 9.1 (1997): 1-42.
[2] Chaudhari, Pratik, et al. "Entropy-sgd: Biasing gradient descent into wide valleys." Journal of Statistical Mechanics: Theory and Experiment 2019.12 (2019): 124018.

Input Smoothing and Weight Smoothing do not Imply Each Other.



Input Smoothing and Weight Smoothing do not Imply Each Other.



- Over-parameterized networks are more robust to adversarial noises in the weight space even when they have the same decision boundary (i.e. the same input smoothness).

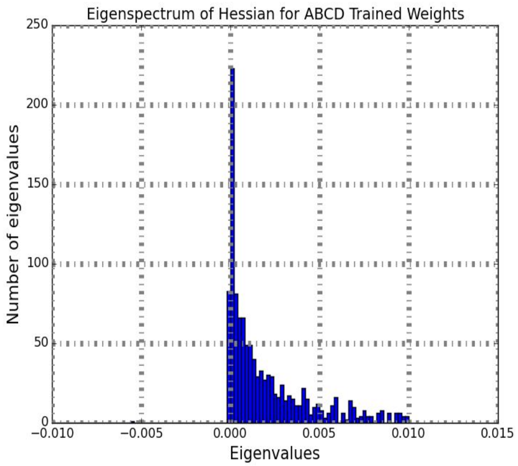
Comparison to State of the art

	Mean Teacher [1]	VAT [2]	Ours
SVHN	3.95	3.86	**3.53** ± 0.24
CIFAR	12.31	10.55	**9.28** ± 0.21

[1] Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.
[2] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2017). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976.

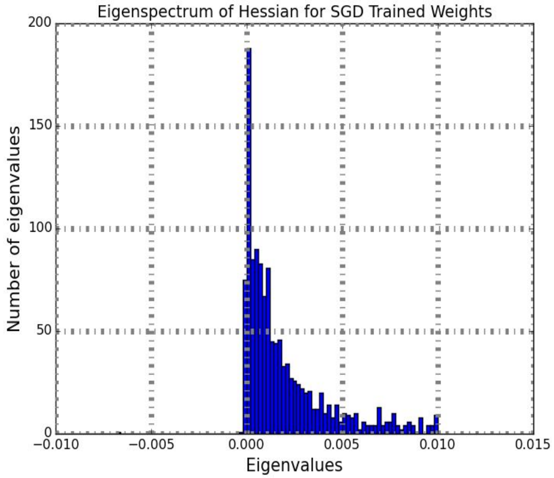
Hessians of the Converged Models

ABCD Trained



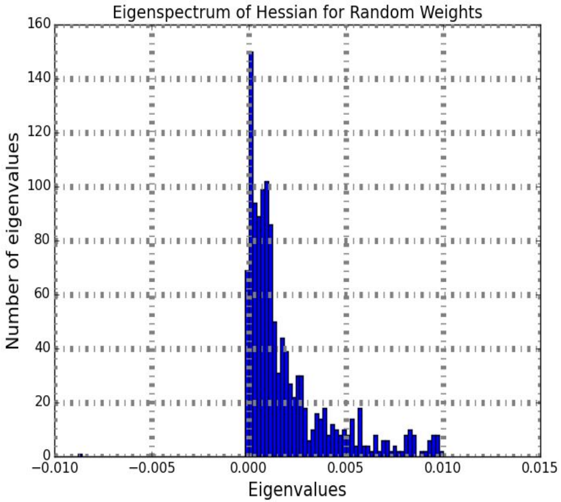
262 almost 0 eigenvalues

SGD Trained



226 almost 0 eigenvalues

Random Weights



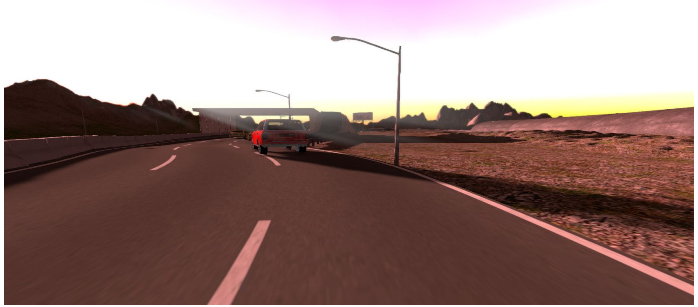
185 almost 0 eigenvalues

Unsupervised Domain Adaptation via Regularized Conditional Alignment

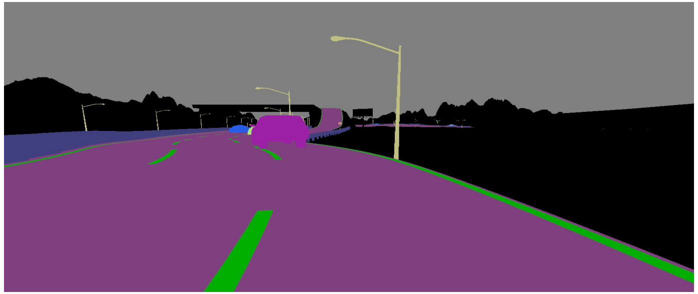
[1] Cicek, Safa, and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019.

Unsupervised Domain Adaptation (UDA)

Synthetic Source



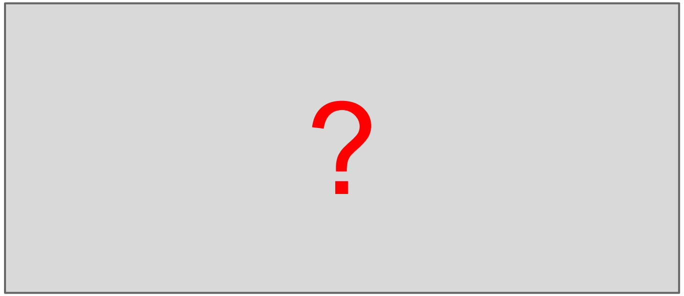
$$(x^s, y^s) \sim P^s$$



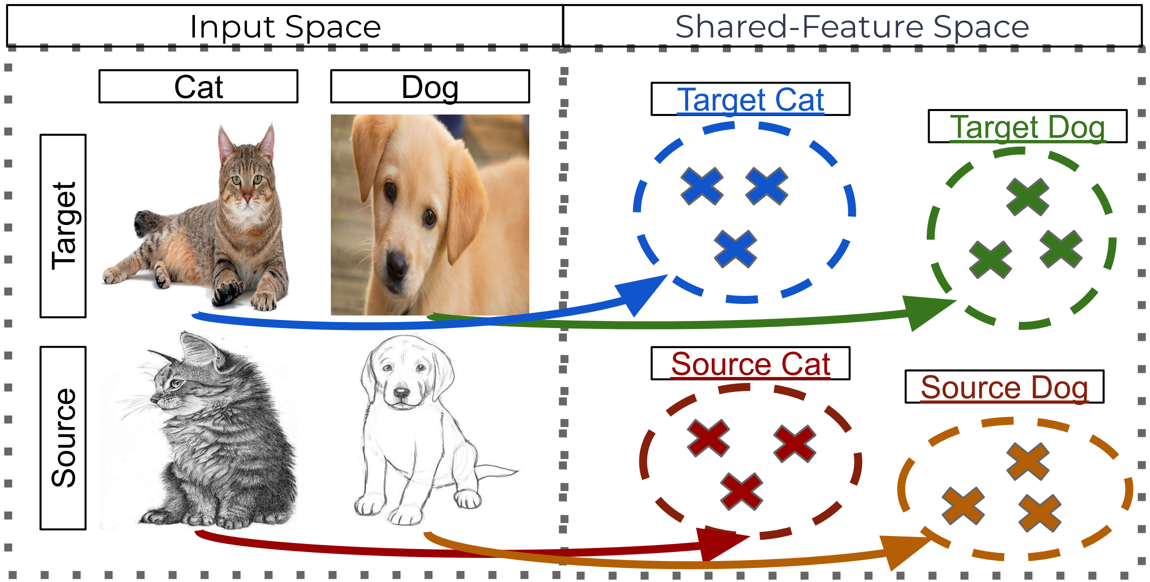
$$KL(P^s || P^t) > 0$$

Real Target

$$x^t \sim P^t_x$$



Shared-Feature Space for UDA

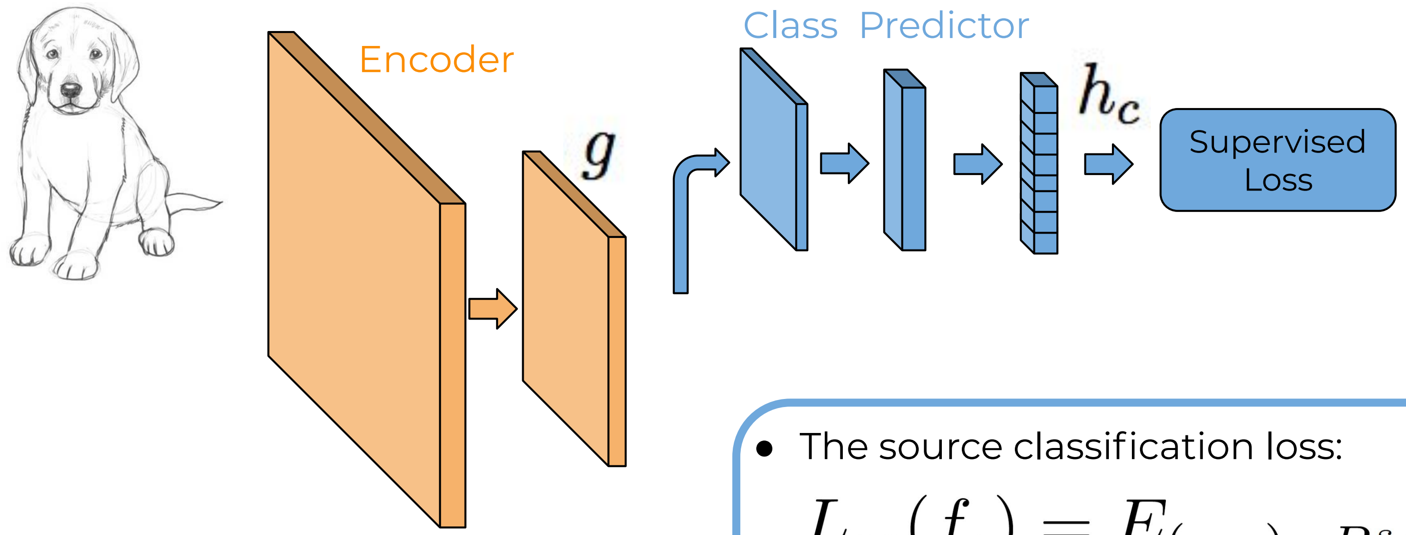


- Moment matching between source and target features (e.g. MMD) [1,2]:

$$||\frac{1}{N^s}\sum_{i=1}^{N^s} g(x_i^s) - \frac{1}{N^t}\sum_{i=1}^{N^t} g(x_i^t)||$$

[1] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.
[2] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791, 2015.

Standard Approach to UDA



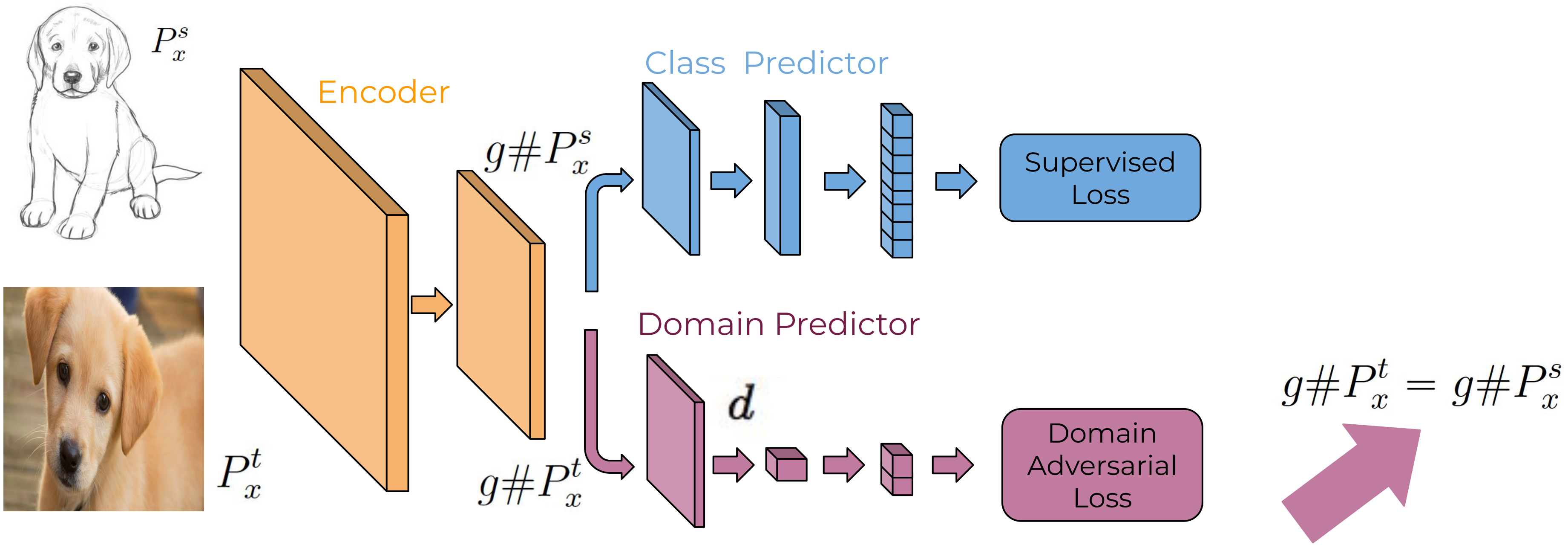
- The source classification loss:

$$L_{sc}(f_c) = E_{(x,y)\sim P^s} \ell_{CE}(f_c(x), y)$$

$$f_c = h_c \circ g$$

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.

Standard Approach to UDA

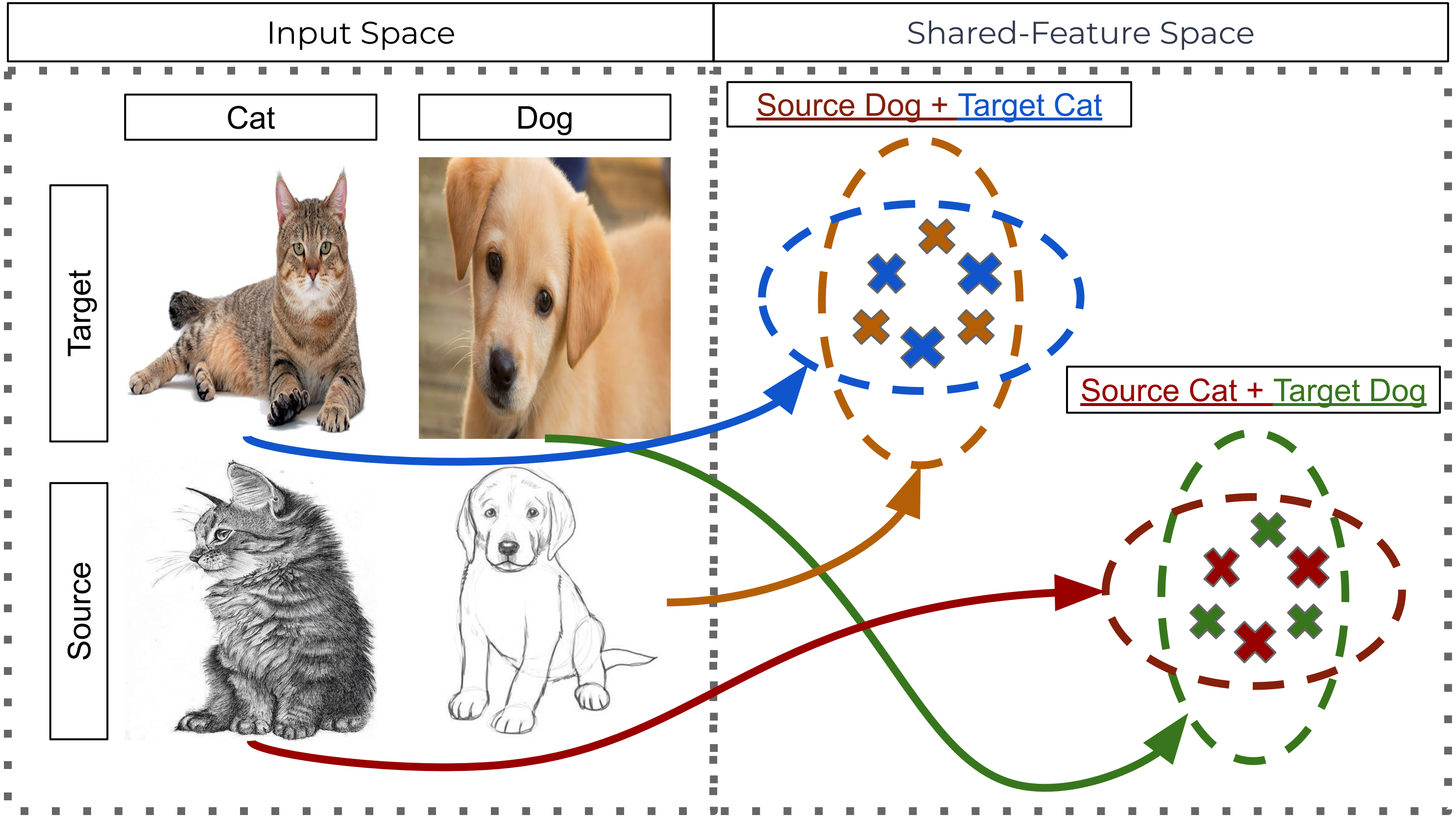


- The domain alignment loss:

$$L_{da}(g,d) = \max_{g} \min_{d} \mathbb{E}_{x\sim P_x^s} \ell_{CE}(d(g(x)), [1, 0]) + \mathbb{E}_{x\sim P_x^t} \ell_{CE}(d(g(x)), [0, 1])$$

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.

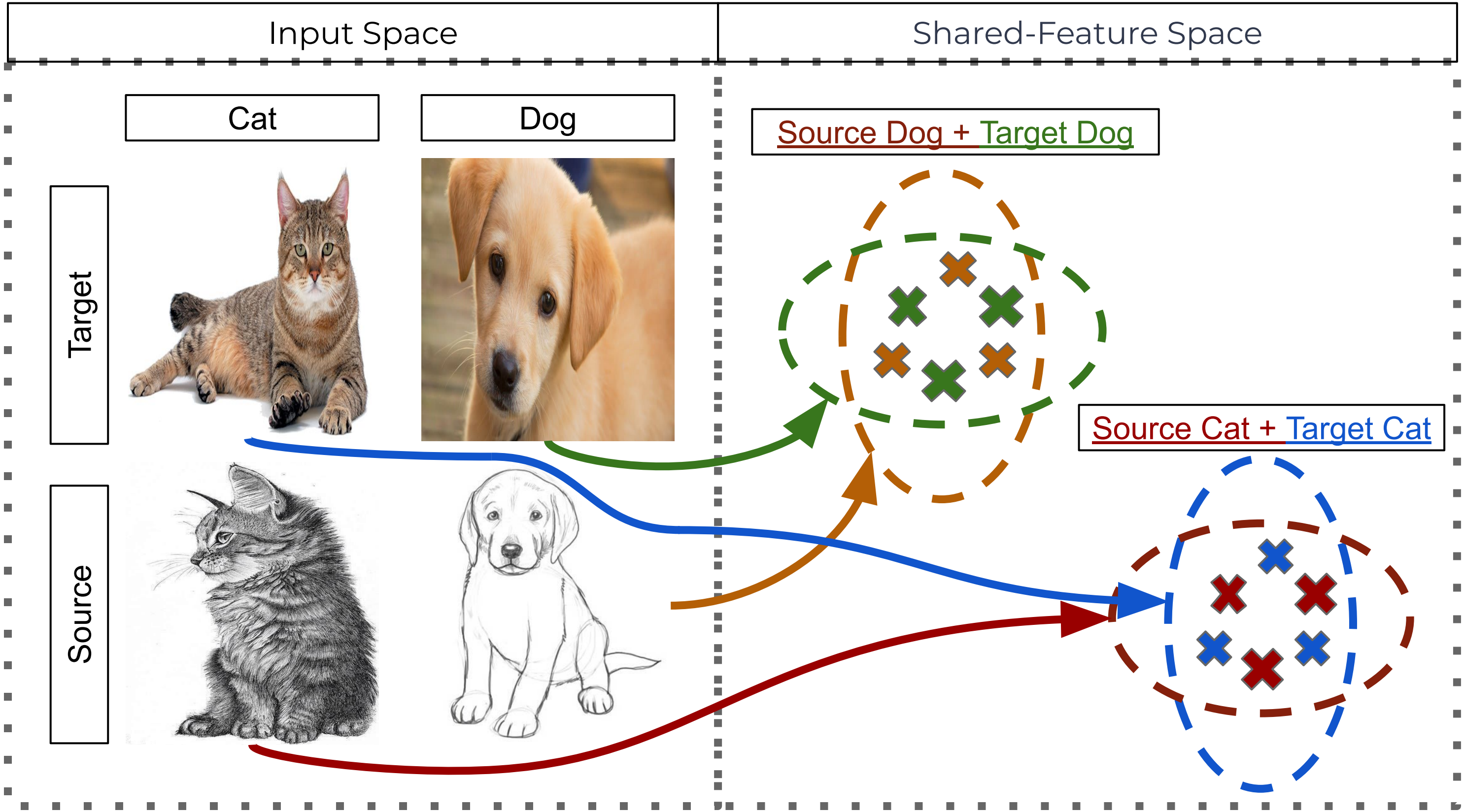
DANN Aligns Marginal Distributions!



- Adversarial domain alignment (e.g. DANN) [1]

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.

Conditional Alignment



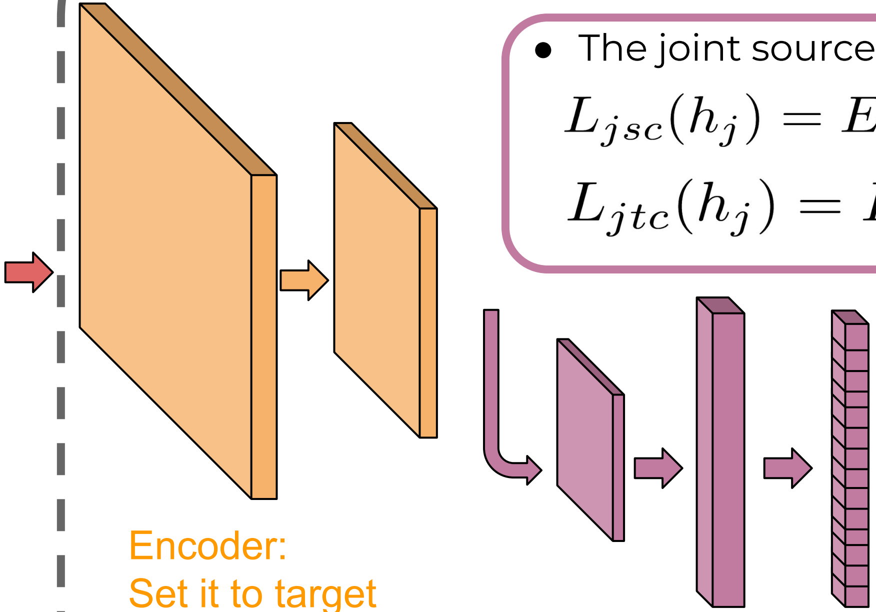
Proposed Method

Conditional Alignment Module

- The joint source and target classification losses:

$$L_{jsc}(h_j) = E_{(x,y)\sim P^s} \ell_{CE}(h_j(g(x)), [y,\ \mathbf{0}])$$

$$L_{jtc}(h_j) = E_{x\sim P^t_x} \ell_{CE}(h_j(g(x)), [\mathbf{0},\ \hat{y}])$$

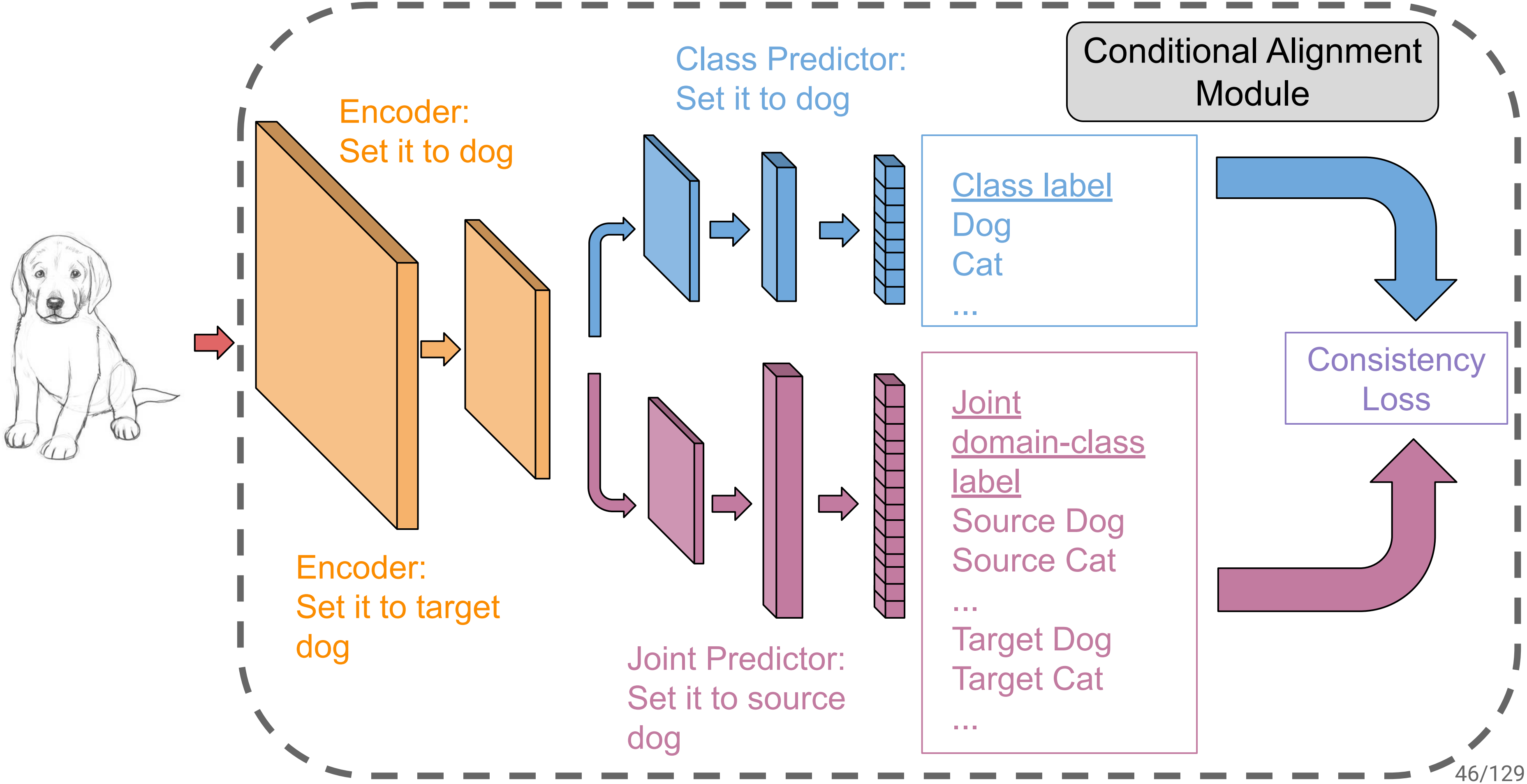


Encoder:
Set it to target dog

Joint Predictor:
Set it to source dog

<u>Joint domain-class label</u>
Source Dog
Source Cat
...
Target Dog
Target Cat
...

Proposed Method



Proposed Method

- Pseudo-labels:

$$\hat{y} = e_k$$

$$k = \arg\max_k f_c(x)[k]$$

- The joint source and target classification losses:

$$L_{jsc}(h_j) = E_{(x,y)\sim P^s} \ell_{CE}(h_j(g(x)), [y, \mathbf{0}])$$

$$L_{jtc}(h_j) = E_{x\sim P_x^t} \ell_{CE}(h_j(g(x)), [\mathbf{0}, \hat{y}])$$

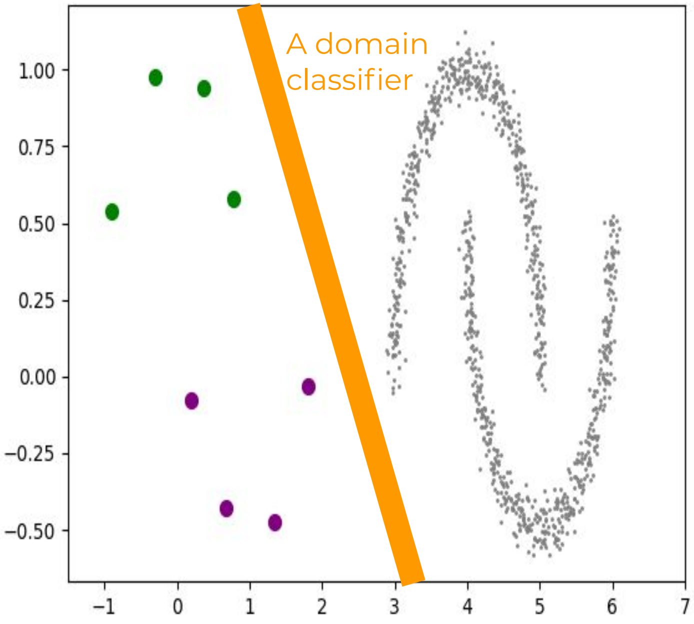
- The joint source and target alignment losses:

$$L_{jsa}(g) = E_{(x,y)\sim P^s} \ell_{CE}(h_j(g(x)), [\mathbf{0}, y])$$

$$L_{jta}(g) = E_{x\sim P_x^t} \ell_{CE}(h_j(g(x)), [\hat{y}, \mathbf{0}])$$

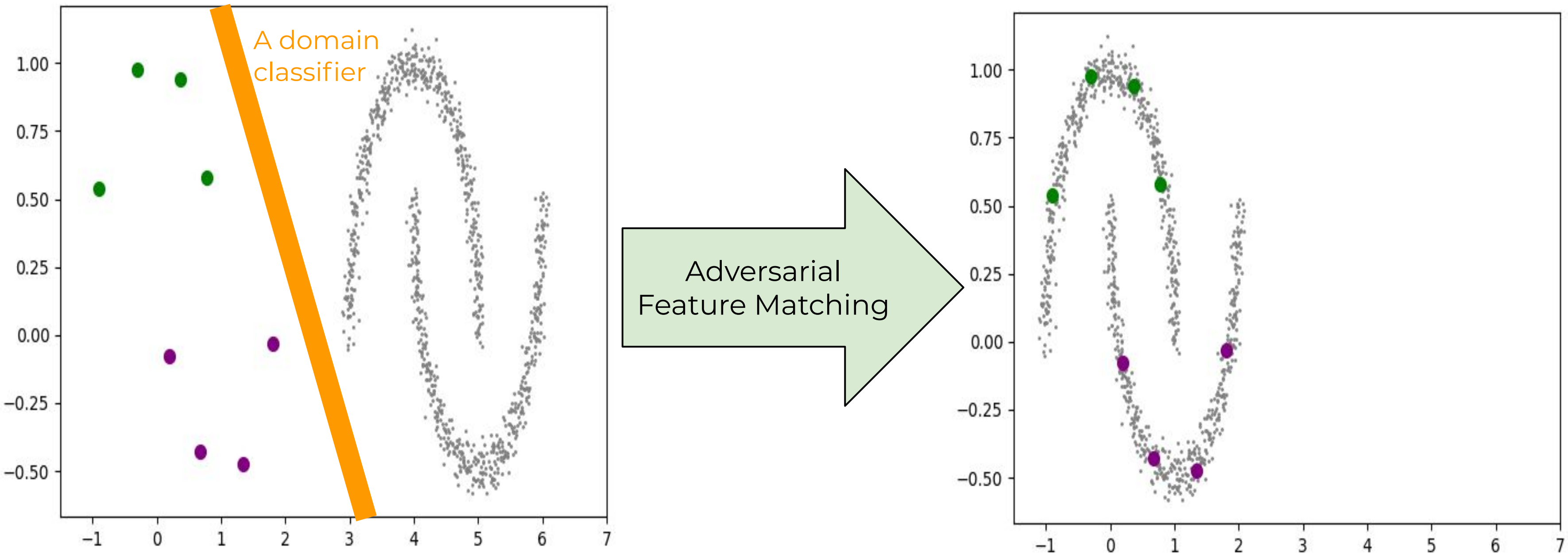
The Joint Discriminator Feedback for Feature Alignment

Exploiting Unlabeled Data with SSL Regularizers



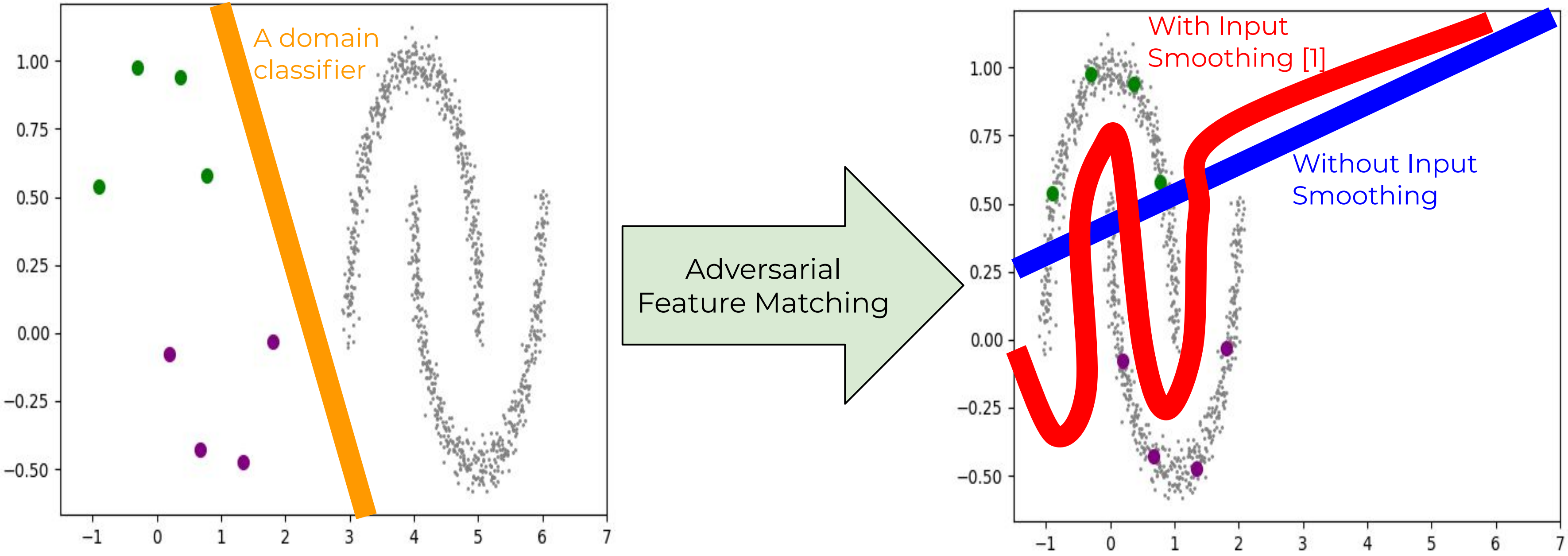
- Gray dots are the learned features for the unlabeled target samples.
- Purple/Green circles are the learned features for the labeled source samples.

Exploiting Unlabeled Data with SSL Regularizers



- Gray dots are the learned features for the unlabeled target samples.
- Purple/Green circles are the learned features for the labeled source samples.

Exploiting Unlabeled Data with SSL Regularizers



- Gray dots are the learned features for the unlabeled target samples.
- Purple/Green circles are the learned features for the labeled source samples.

[1] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976, 2017.

Analysis

Proposition 1. *The optimal joint predictor h_j minimizing $L_{jsc}(h_j) + L_{jtc}(h_j)$ for any feature z with non-zero measure either on $g\#P_x^s(z)$ or $g\#P_x^t(z)$ is*

$$h_j(z)[i] = \frac{g\#P^s(z, y = e_i)}{g\#P_x^s(z) + g\#P_x^t(z)}$$

$$h_j(z)[i+K] = \frac{g\#P^t(z, y = e_i)}{g\#P_x^s(z) + g\#P_x^t(z)} \textit{ for } i \in \{1, ..., K\}$$

Theorem 1. *The objective $L_{jsa}(g) + L_{jta}(g)$ is minimized for the given optimal joint predictor if and only if*

$$\boxed{g\#P^s(z|y = e_k) = g\#P^t(z|y = e_k)}$$

$$g\#P^s(z|y = e_k) > 0 \Rightarrow g\#P^s(z|y = e_i) = 0 \;\; \textit{for } i \neq k \textit{ for any } y = e_k \textit{ and } z.$$

Comparison to SOA UDA Methods

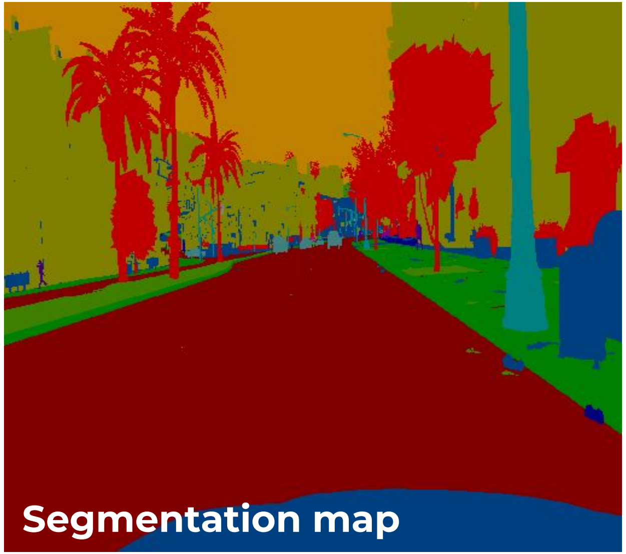
Source dataset	MNIST	SVHN	CIFAR	STL	SYN-DIGITS	MNIST
Target dataset	SVHN	MNIST	STL	CIFAR	SVHN	MNIST-M
DANN [1]	60.6	68.3	78.1	62.7	90.1	94.6
VADA + IN [2]	73.3	94.5	78.3	71.4	94.9	95.7
Ours	**89.19**	**99.33**	**81.65**	**77.76**	**96.22**	**99.47**
Source-only	44.21	70.58	79.41	65.44	85.83	70.28
Target-only	94.82	99.28	77.02	92.04	96.56	99.87

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.
[2] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:1802.08735, 2018.

Disentangled Image Generation for Unsupervised Domain Adaptation

[1] Cicek, Safa, Zhaowen Wang, Hailin Jin, Stefano Soatto, Generative Feature Disentangling for Unsupervised Domain Adaptation. Proceedings of the European Conference on Computer Vision (ECCV) Workshops. (2020).

Image Translation Approach



y

$x \sim P^g(x|y, d=0)$

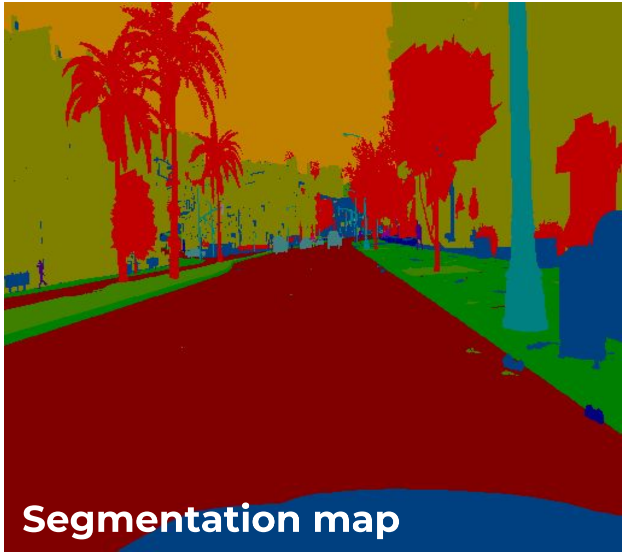
$x \sim P^g(x|y, d=1)$

- We generate the images using GauGAN [1].

[1] Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Image Translation Approach

In reality, Cityscapes (Germany) do not have palm trees 😁



y

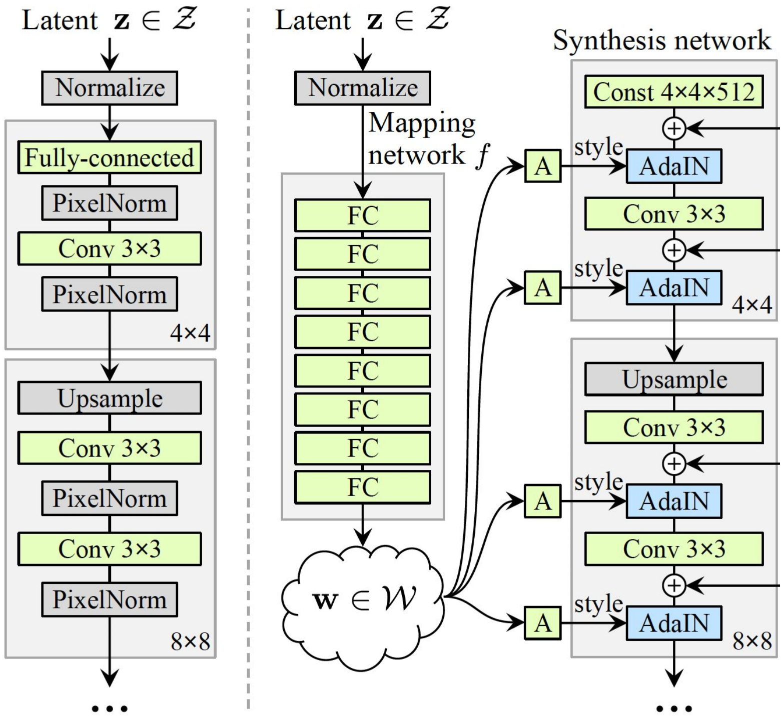
$x \sim P^g(x|y, d = 1)$

- We generate the images using GauGAN [1].

[1] Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

StyleGAN

$$z \sim N(0, I)$$



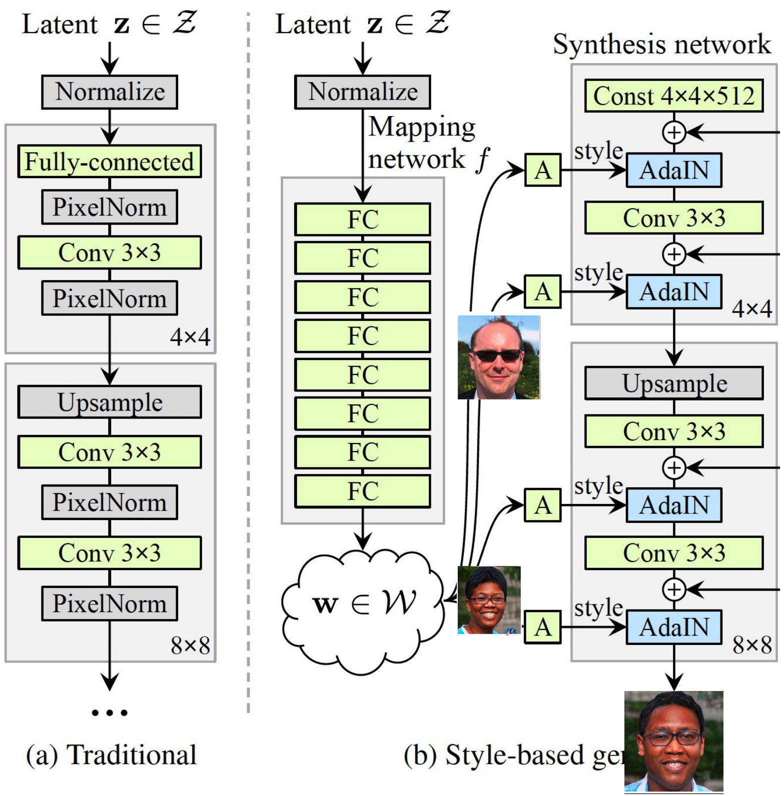
(a) Traditional

(b) Style-based generator

[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

StyleGAN

$$z \sim N(0, I)$$

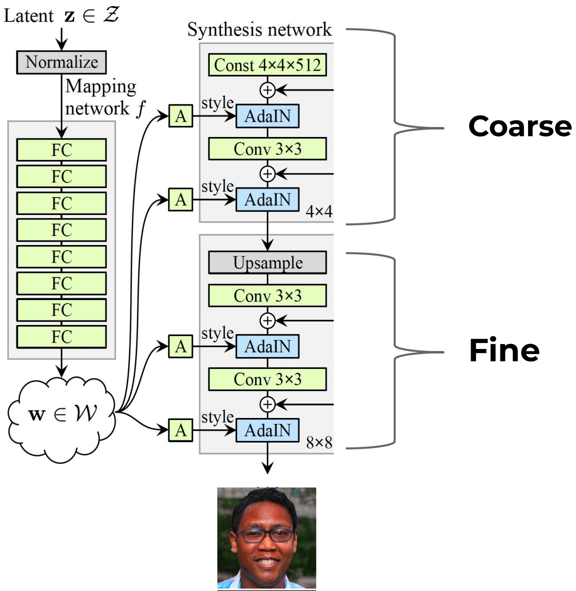


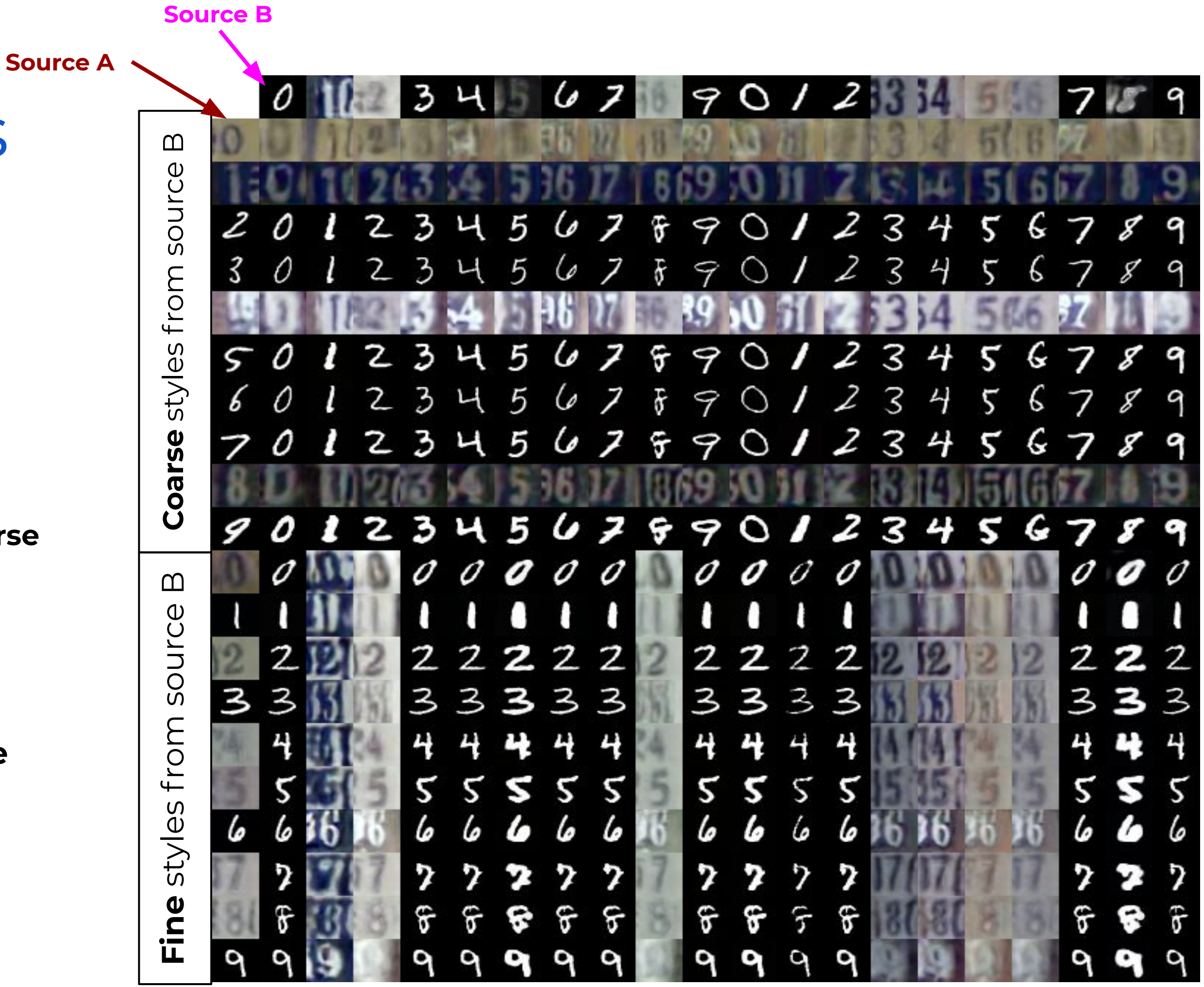
(a) Traditional (b) Style-based ge

Style Mixing

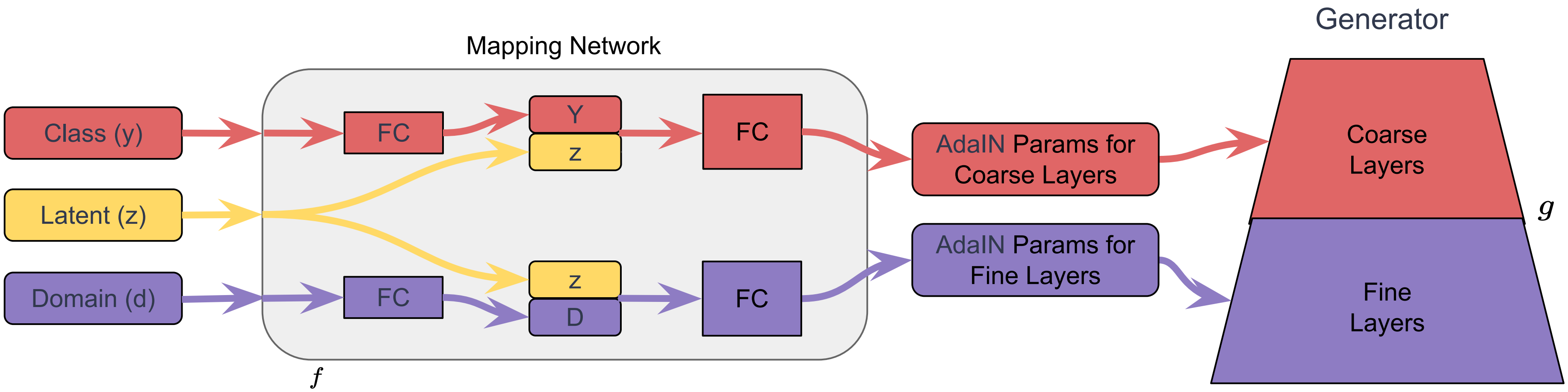
[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Mixing Learned Styles for Multiple Domains

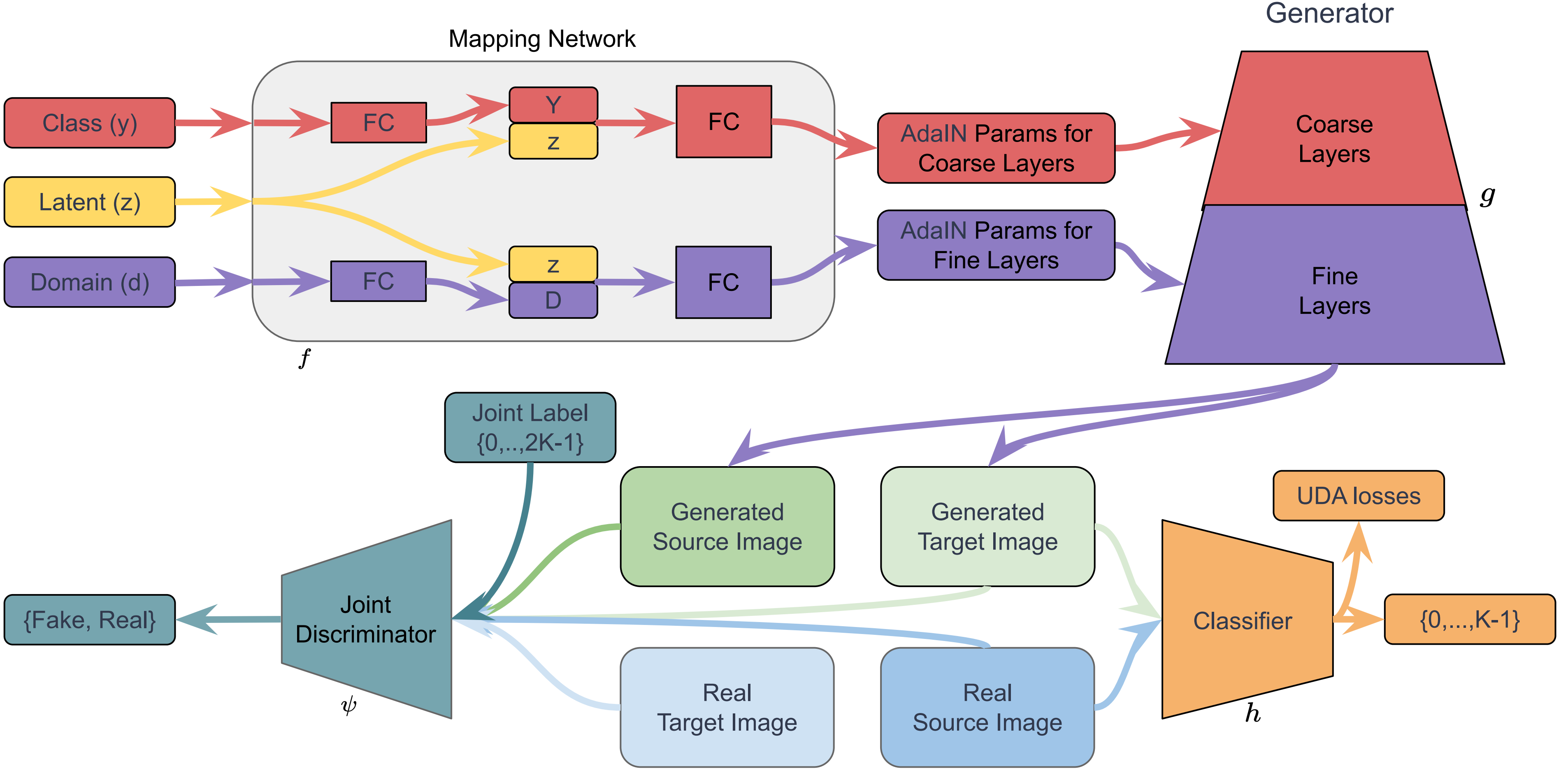




Explicit Regularization for UDA



Explicit Regularization for UDA



Colored Background and Colored Digit Datasets

src

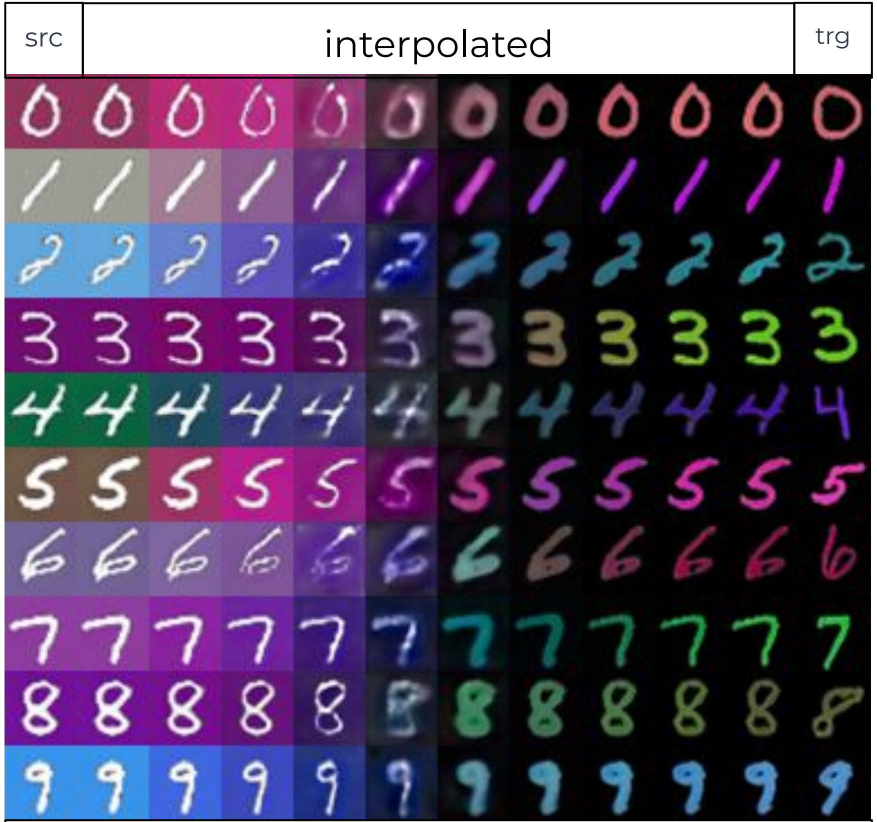
Colored background

Colored digit

trg

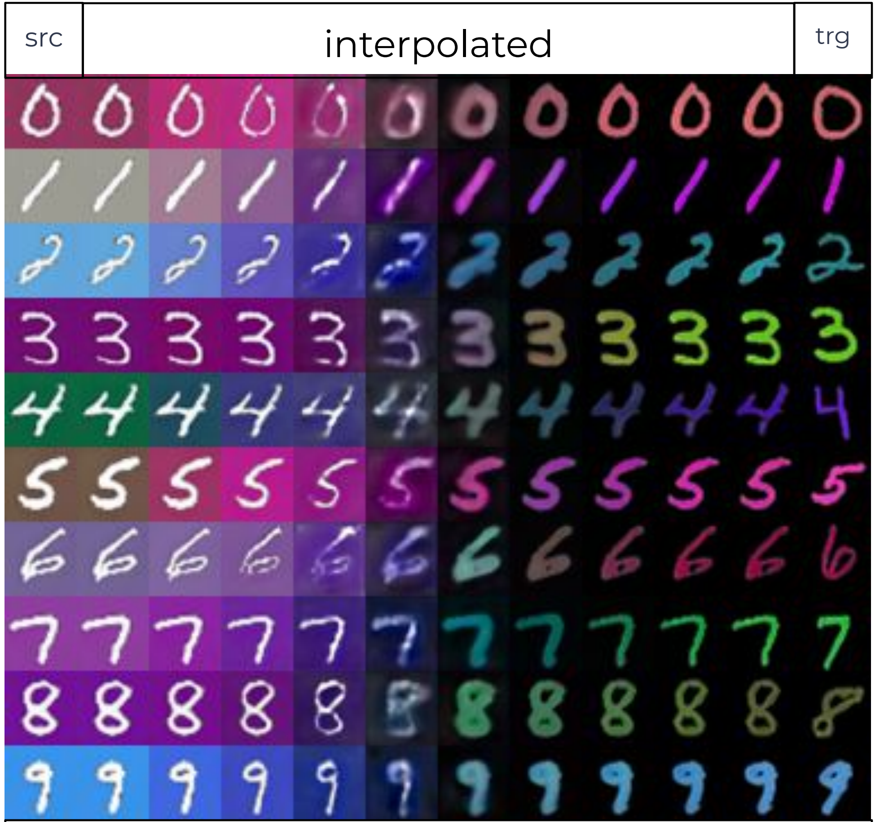
[1] Gonzalez-Garcia, Abel, Joost Van De Weijer, and Yoshua Bengio. "Image-to-image translation for cross-domain disentanglement." *Advances in neural information processing systems*. 2018.

Interpolation of the ***Fine*** Layer Parameters

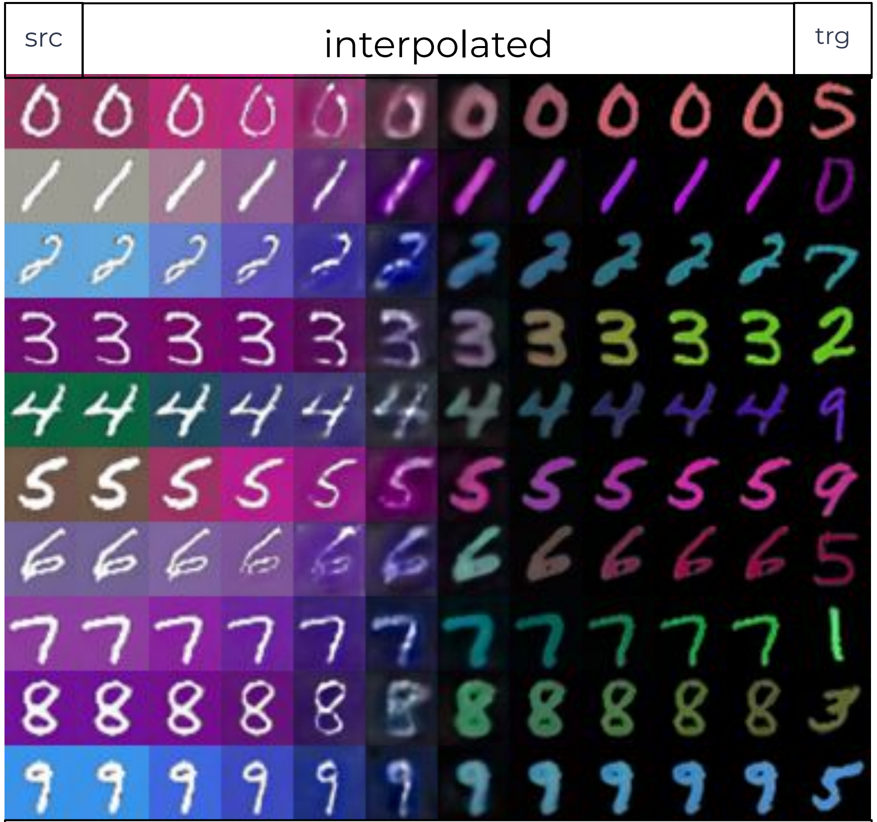


Generated source and target images have the same class label.

Interpolation of the *Fine* Layer Parameters

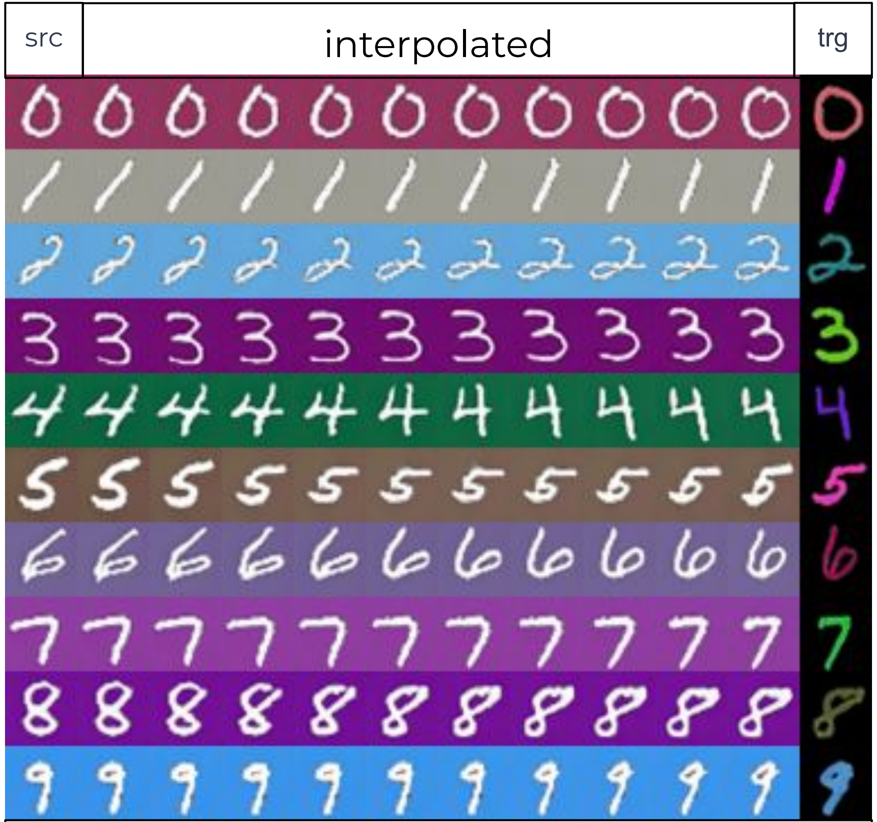


Generated source and target images have the **same** class label.



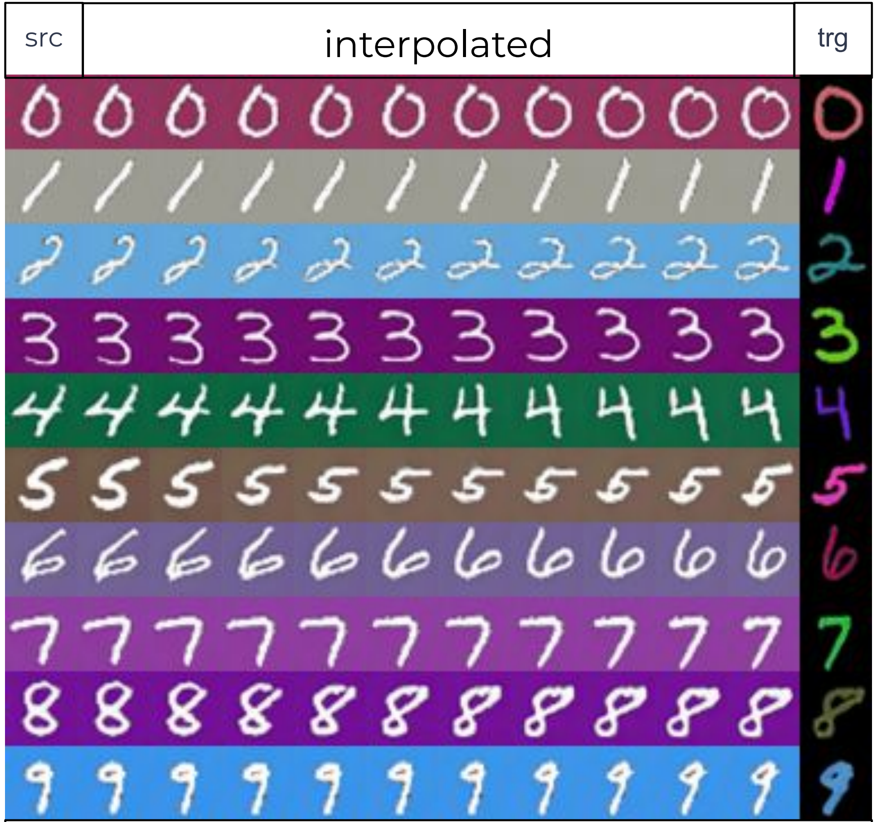
Generated source and target images have **different** class labels.

Interpolation of the ***Coarse*** Layer Parameters

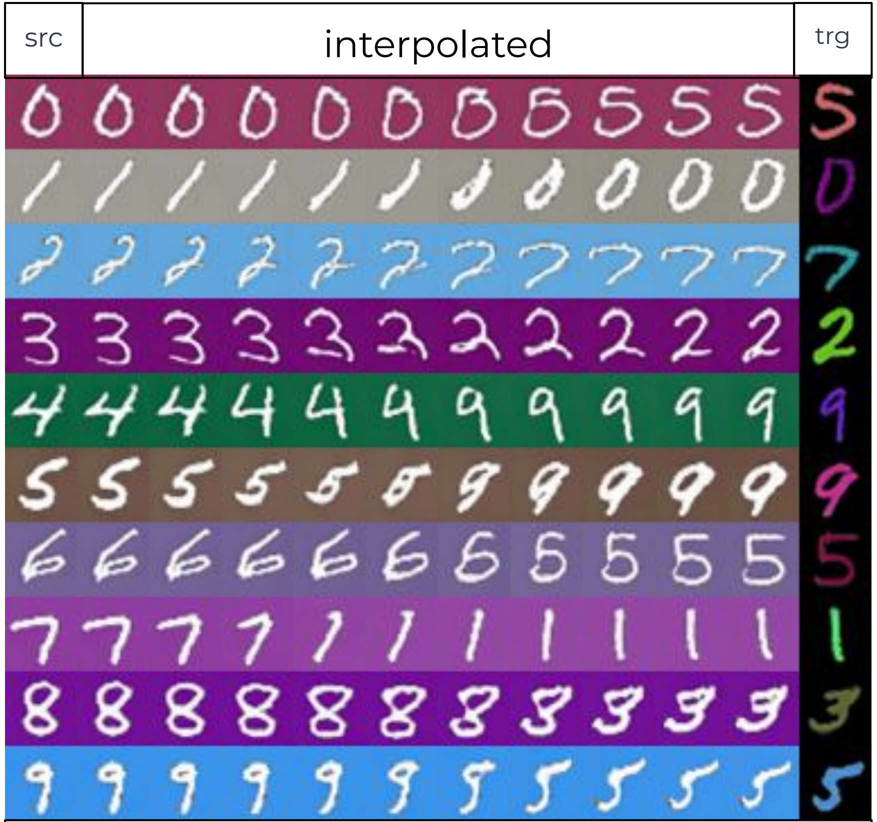


Generated source and target images have the same class label.

Interpolation of the ***Coarse*** Layer Parameters

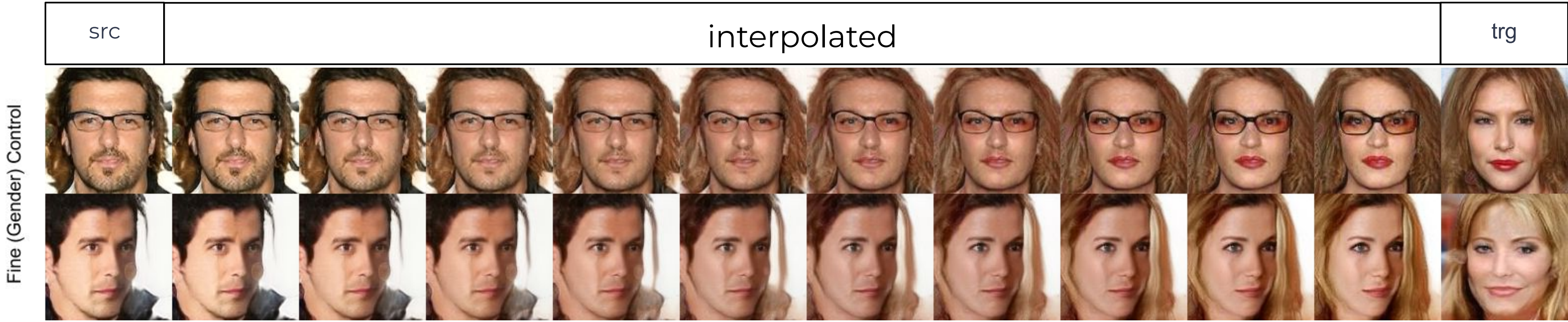


Generated source and target images have the **same** class label.

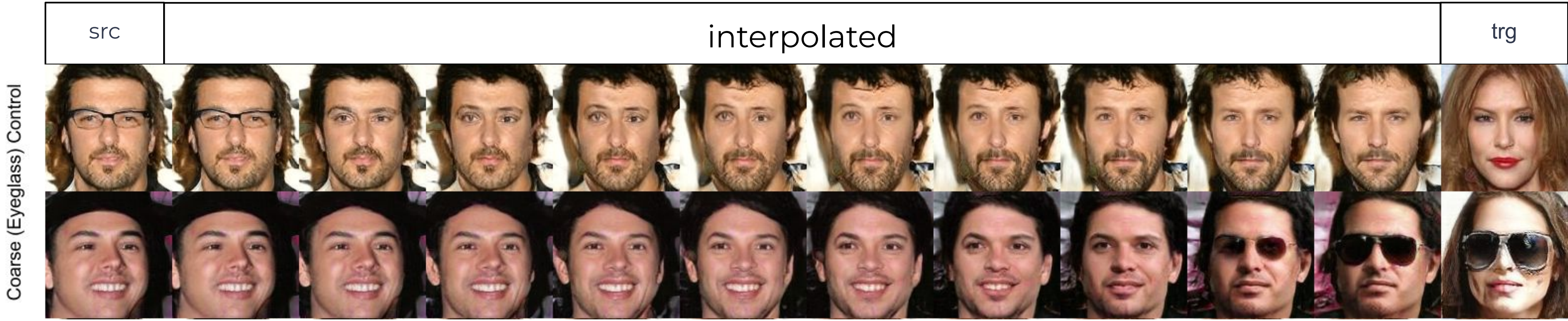


Generated source and target images have **different** class labels.

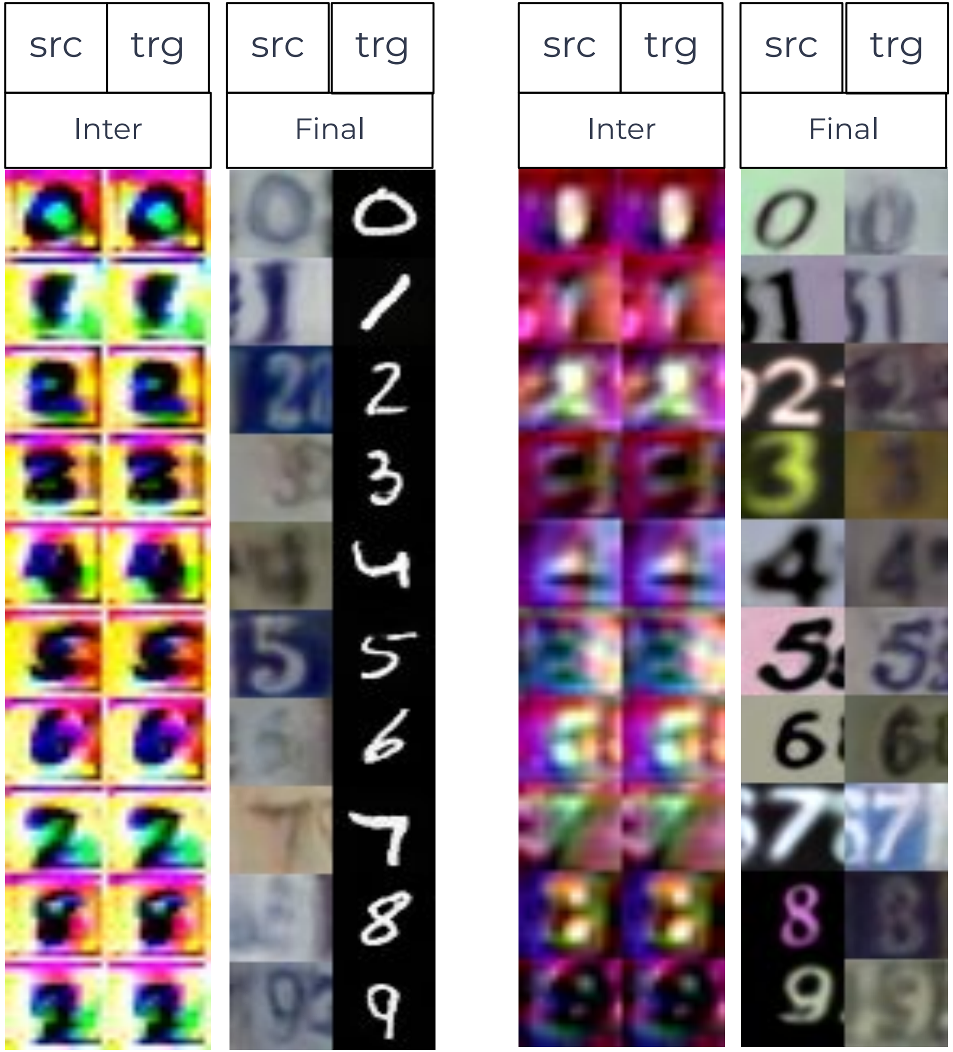
Interpolation of the ***Fine*** Layer Parameters



Interpolation of the ***Coarse*** Layer Parameters



Learned Shared Representations at the Intermediate Layers:



Results in MSDA Benchmarks

SYN-DIGITS, MNIST, USPS, SVHN -> MNIST-M

SYN-DIGITS, MNIST, USPS, MNIST-M -> SVHN

SVHN, MNIST, USPS, MNIST-M -> SYN-DIGITS

Results in MSDA Benchmarks

SYN-DIGITS, MNIST, USPS, SVHN -> MNIST-M

SYN-DIGITS, MNIST, USPS, MNIST-M -> SVHN

SVHN, MNIST, USPS, MNIST-M -> SYN-DIGITS

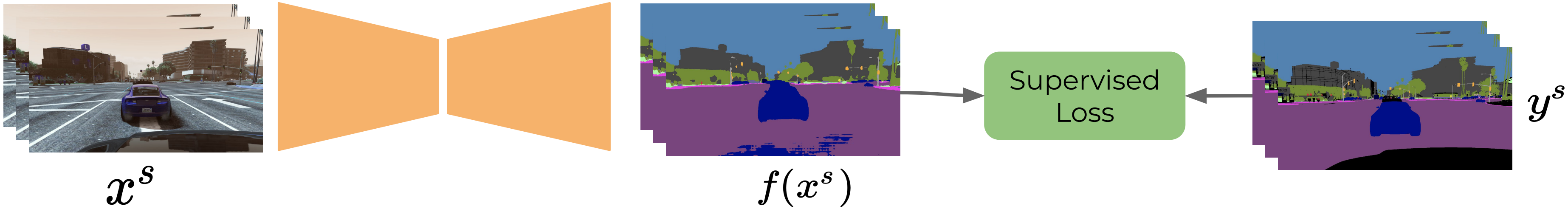
Target dataset	SVHN	SYN-DIGITS	MNIST	USPS	MNIST-M
DCTN [1]	77.5	NR	NR	NR	70.9
M^3SDA [2]	81.32	89.58	98.58	96.14	72.82
Ours	**90.71**	**98.91**	**99.65**	**97.20**	**98.45**

[1] Xu, Ruijia, et al. "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[2] Peng, Xingchao, et al. "Moment matching for multi-source domain adaptation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

Spatial Class Distribution Shift in Unsupervised Domain Adaptation: Local Alignment Comes to Rescue

[1] Cicek, Safa, Ning Xu, Zhaowen Wang, Hailin Jin, Stefano Soatto, Spatial Class Distribution Shift in Unsupervised Domain Adaptation. Asian Conference on Computer Vision (ACCV). 2020.

Standard Approach to UDA

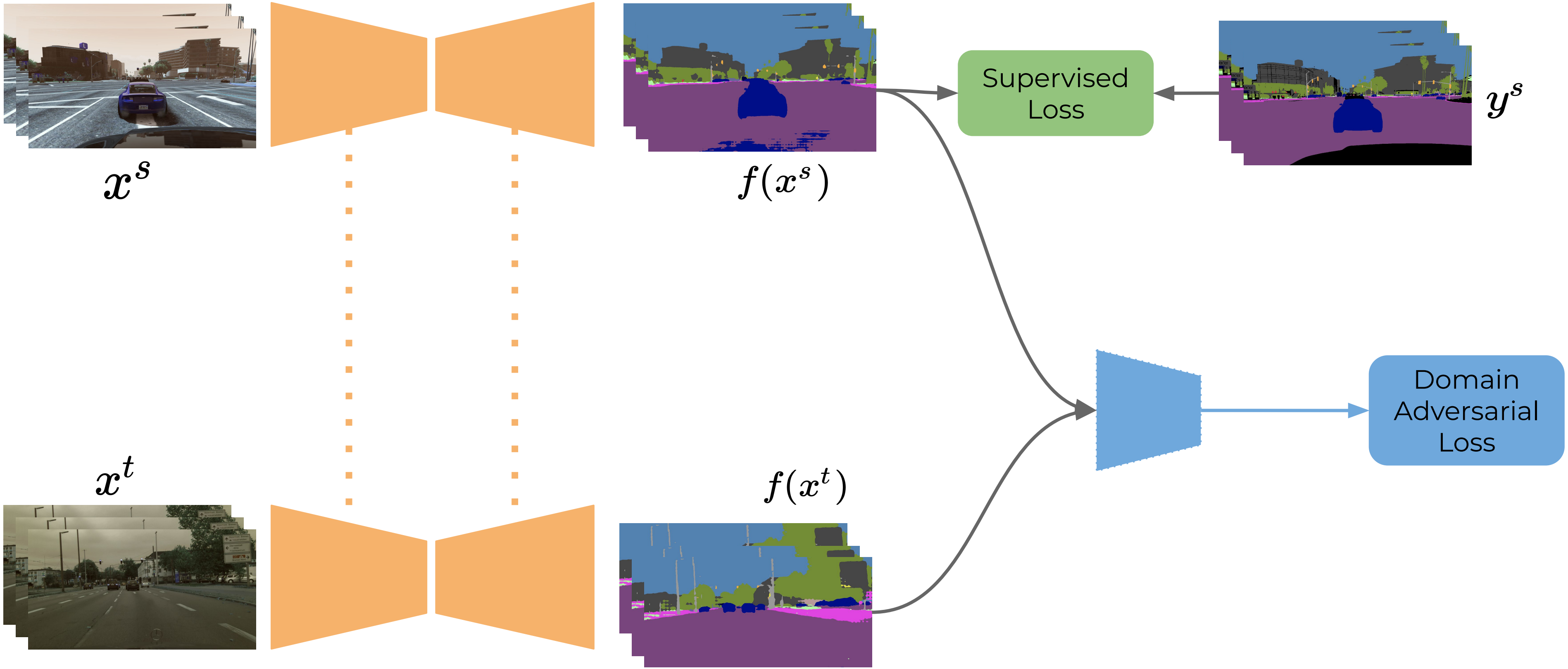


- Source classification loss

$$L_{ce}(P^s; f) := \mathbb{E}_{(x^s, y^s) \sim P^s} \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \ell_{CE}(f(x^s)_{ij}; y^s_{ij})$$

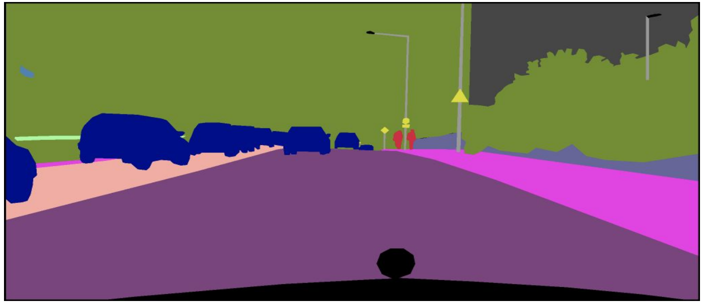
[1] Vu, Tuan-Hung, et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

Standard Approach to UDA

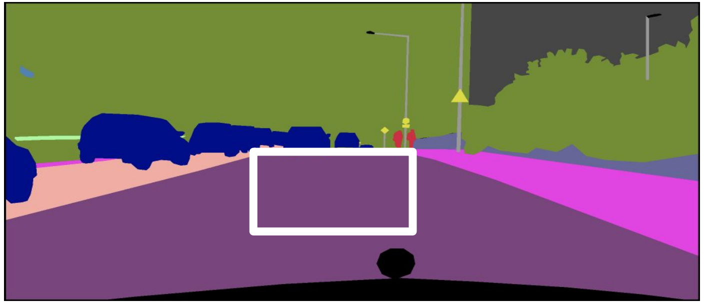


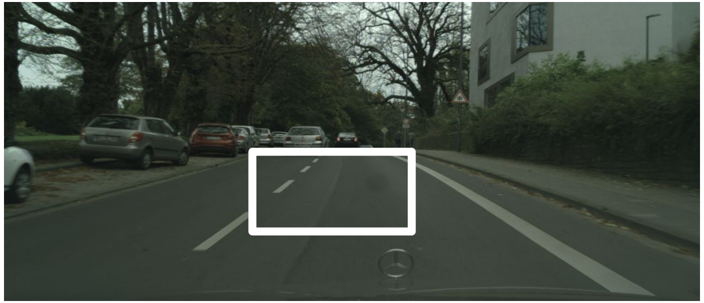
[1] Vu, Tuan-Hung, et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

Spatial-class-distribution Shift

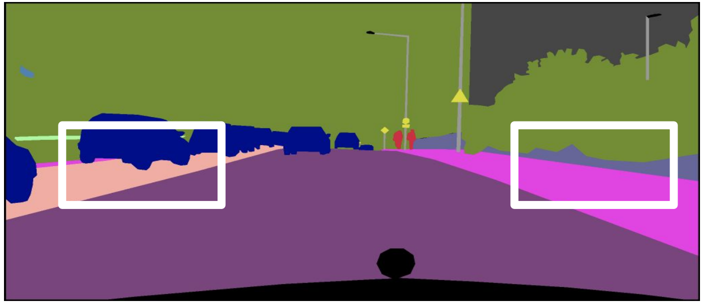


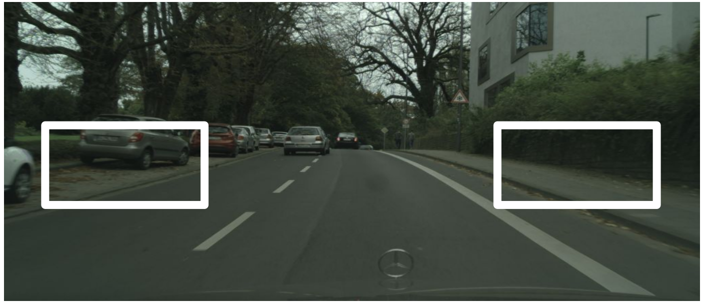
Spatial-class-distribution Shift



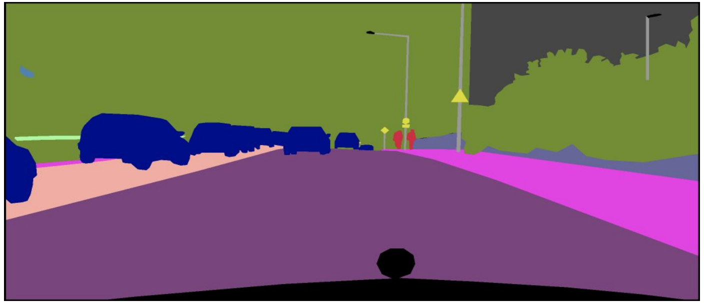


Spatial-class-distribution Shift

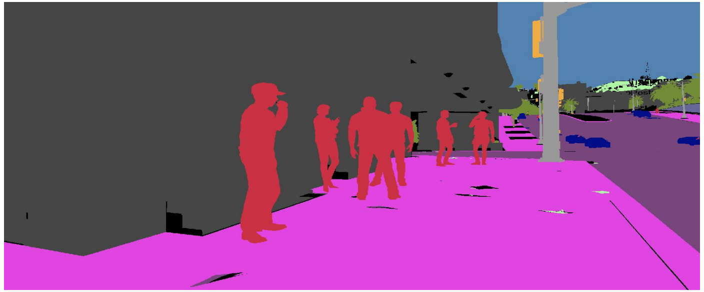




Spatial-class-distribution Shift

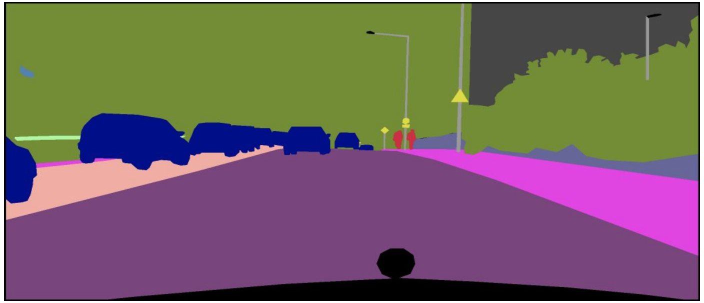


Domain-I (Cityscapes): Images are captured from dashcam view and scenarios are realistic.

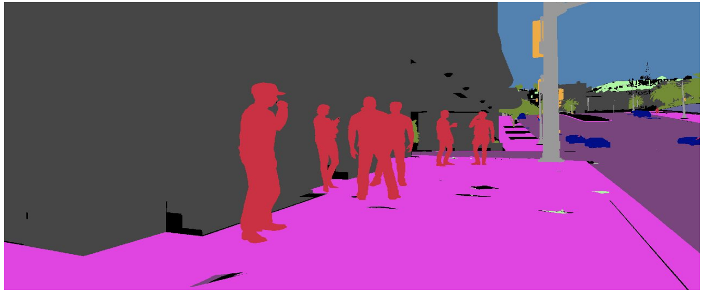


Domain-II (GTA5): Images are captured in unrealistic scenarios e.g. vehicle driving on the sidewalk.

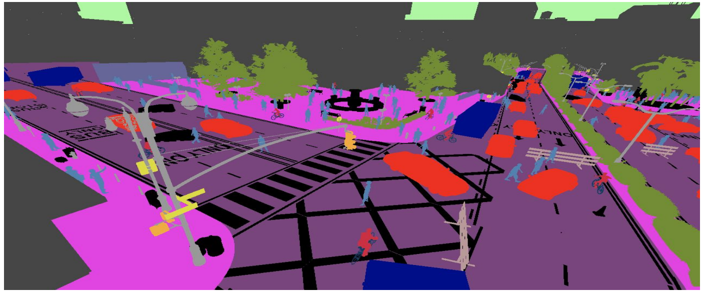
Spatial-class-distribution Shift




Domain-I (Cityscapes): Images are captured from dashcam view and scenarios are realistic.

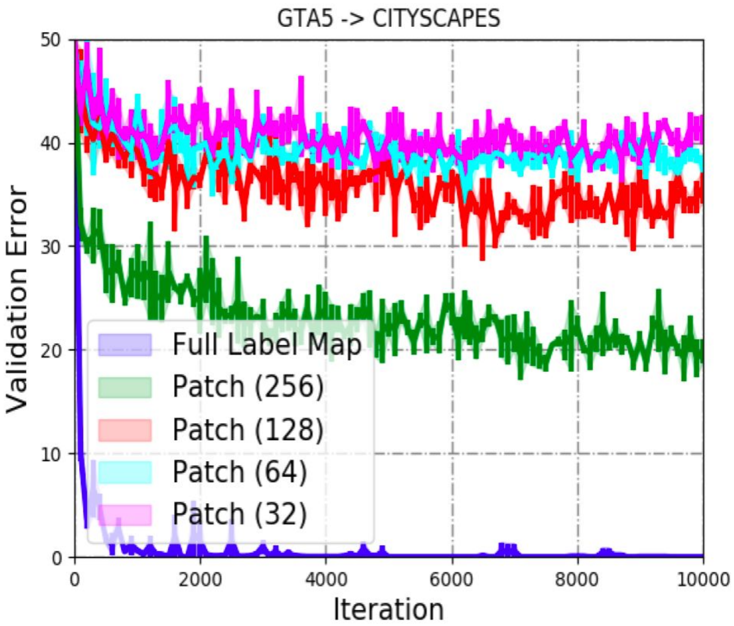


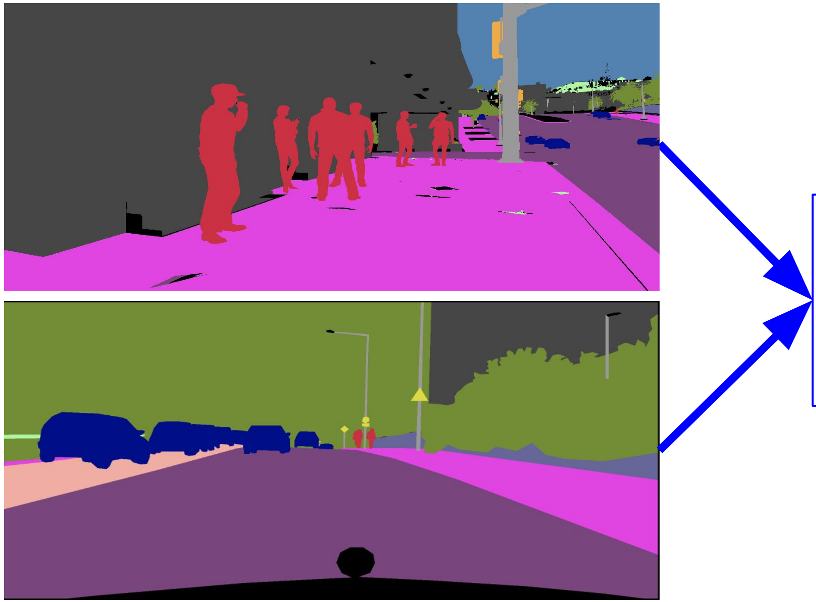

Domain-II (GTA5): Images are captured in unrealistic scenarios e.g. vehicle driving on the sidewalk.




Domain-III (SYNTHIA): Images are captured with random camera views.

Spatial-class-distribution shift correlates with the receptive field.

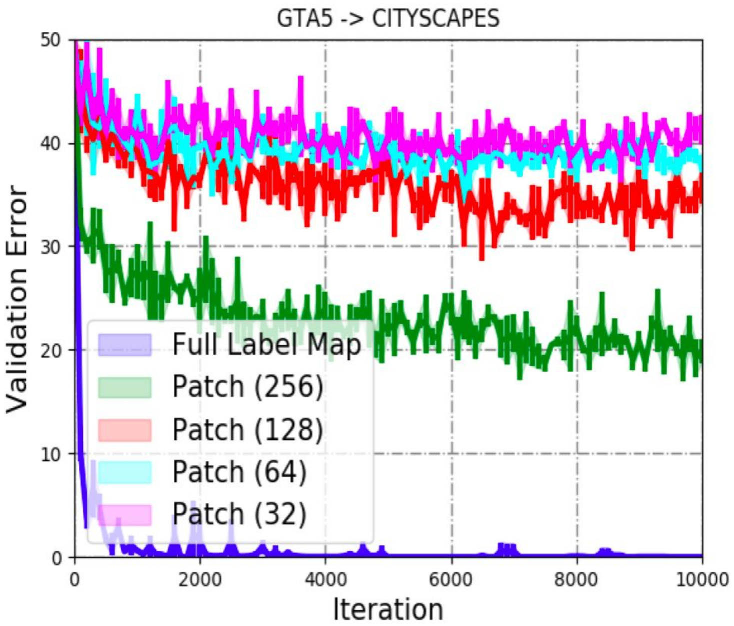


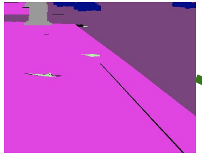


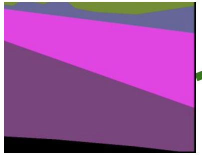
What is the domain: GTA5 or Cityscapes?

- Validation errors for a binary classifier trained to distinguish binary domain labels from **segmentation** maps.

Spatial-class-distribution shift correlates with the receptive field.



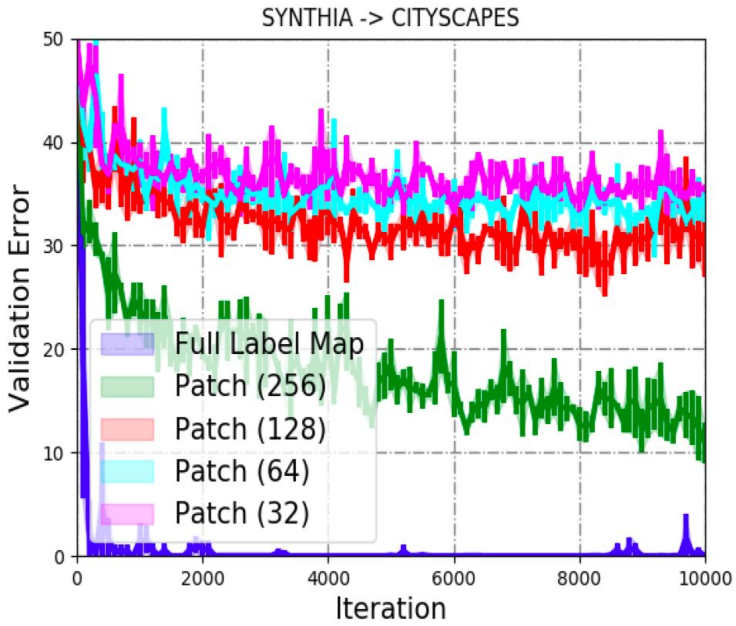


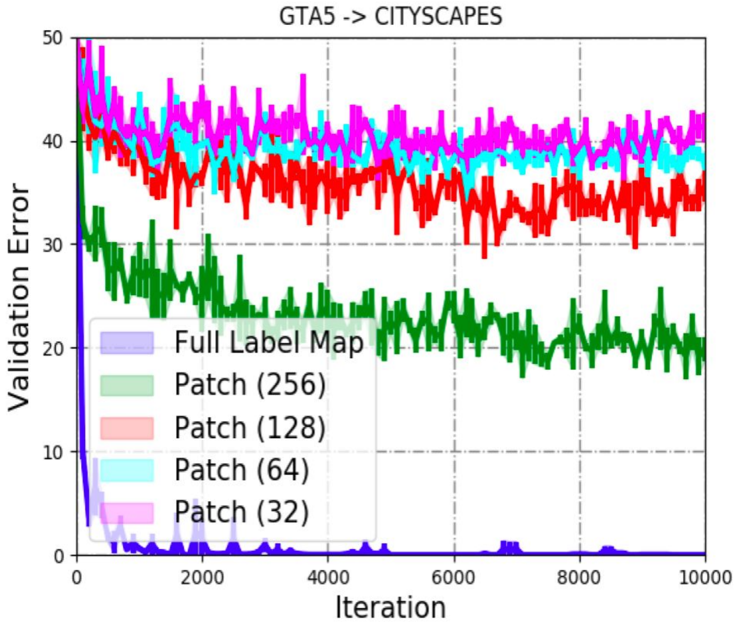


What is the domain: GTA5 or Cityscapes?

- Domain is less identifiable for smaller receptive fields.

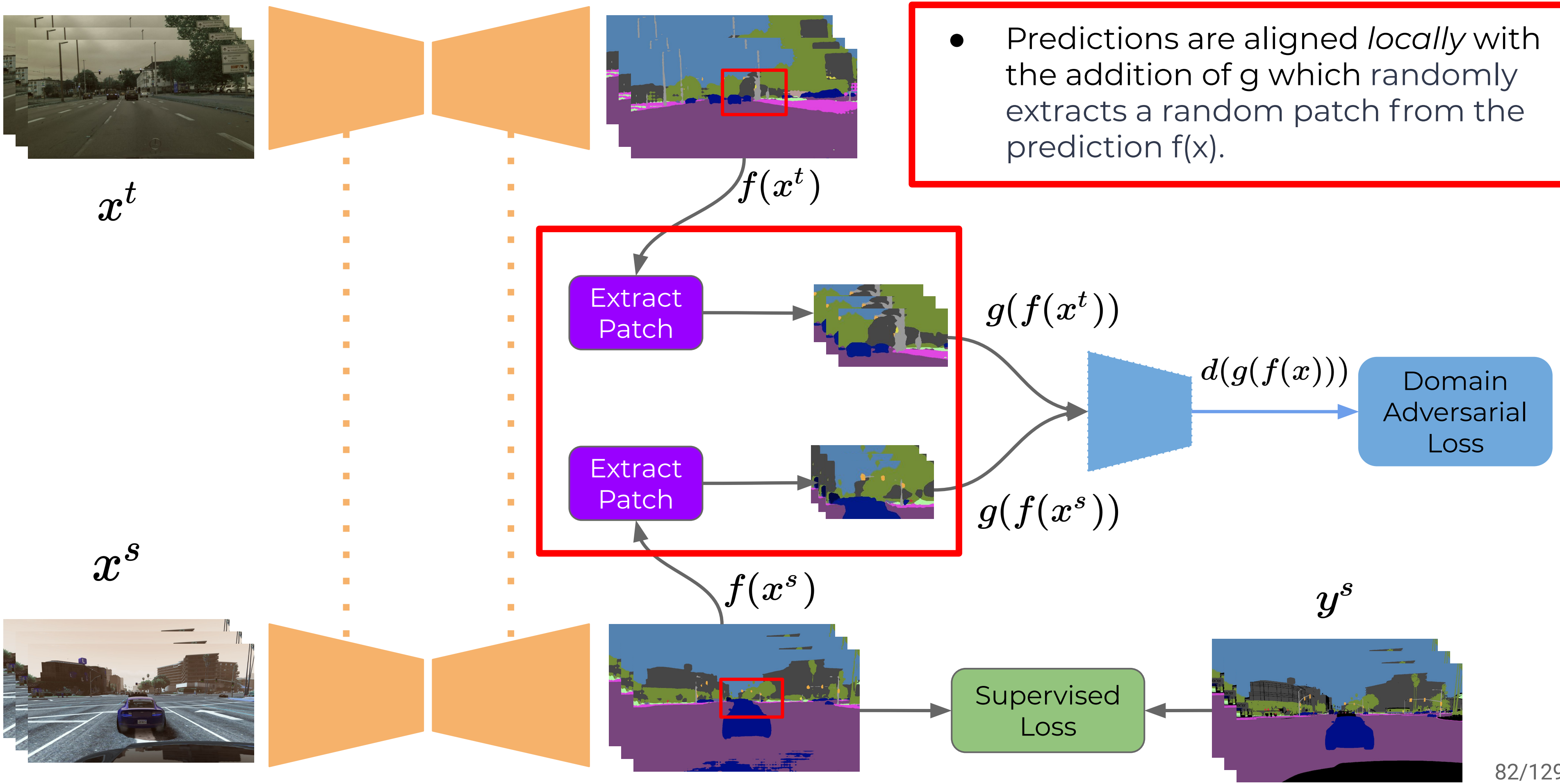
Spatial-class-distribution shift correlates with the receptive field.





- Errors for SYNTHIA are slightly lower due to the larger spatial-class shift between SYNTHIA and Cityscapes.

Proposed Method



- Predictions are aligned *locally* with the addition of g which randomly extracts a random patch from the prediction f(x).

Objective Functions

- Adversarial domain alignment loss from [1]:

$$L_{advent}(P_x^s, P_x^t; f, d) := \mathbb{E}_{x^s \sim P_x^s, x^t \sim P_x^t} \ell_{CE}\left(\overline{\psi}(x^s), [0,1]\right) + \ell_{CE}\left(\overline{\psi}(x^t), [1,0]\right)$$

$$\min_f \max_d L_{ce}(P^s; f) - \lambda L_{advent}(P_x^s, P_x^t; f, d)$$

where $\overline{\psi}(x) := \boxed{d(g(h(f(x))))}$ $\quad h(y_{kij}) = -y_{kij} \log y_{kij} \quad d : x \mapsto \mathbb{R}^2$

$\boxed{g \text{ randomly extracts a patch of size } i < H \text{ and } j < W}$

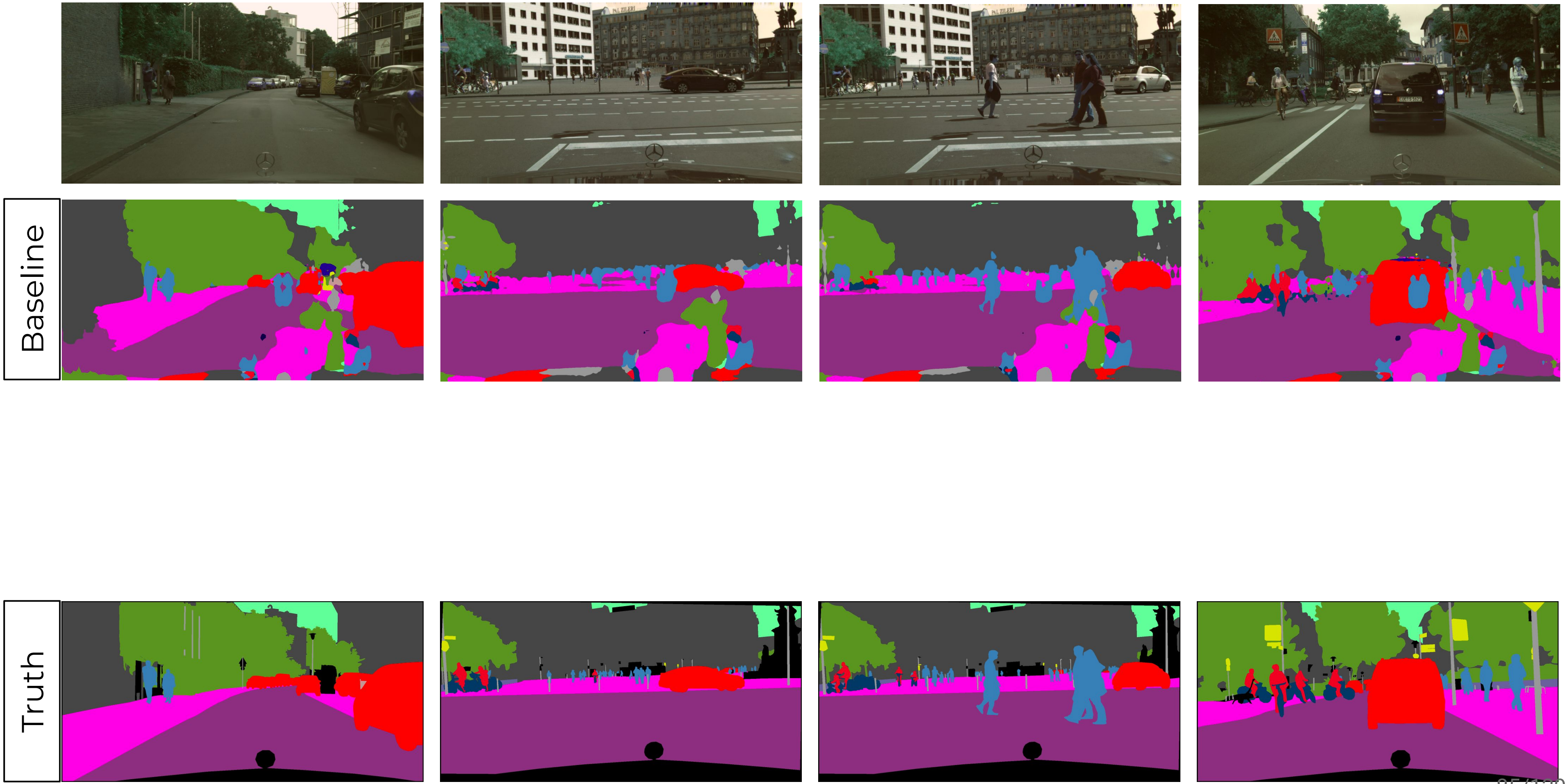
[1] Vu, Tuan-Hung, et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Quantitative Results: Comparison to SOA

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.
AdvEnt [1]	87.0	44.1	79.7	9.6	0.6	24.3	4.8	7.2	80.1
A+E [1]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4
MRKLD[2]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8
Ours	90.6	51.34	81.96	11.77	0.32	29.51	11.72	12.38	82.69
Method	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU	mIoU-13
AdvEnt [1]	83.6	56.4	23.7	72.7	32.6	12.8	33.7	40.8	47.6
A+E [1]	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
MRKLD[2]	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
Ours	84.7	58.57	24.73	81.94	36.37	17.11	41.75	**44.84**	**51.99**

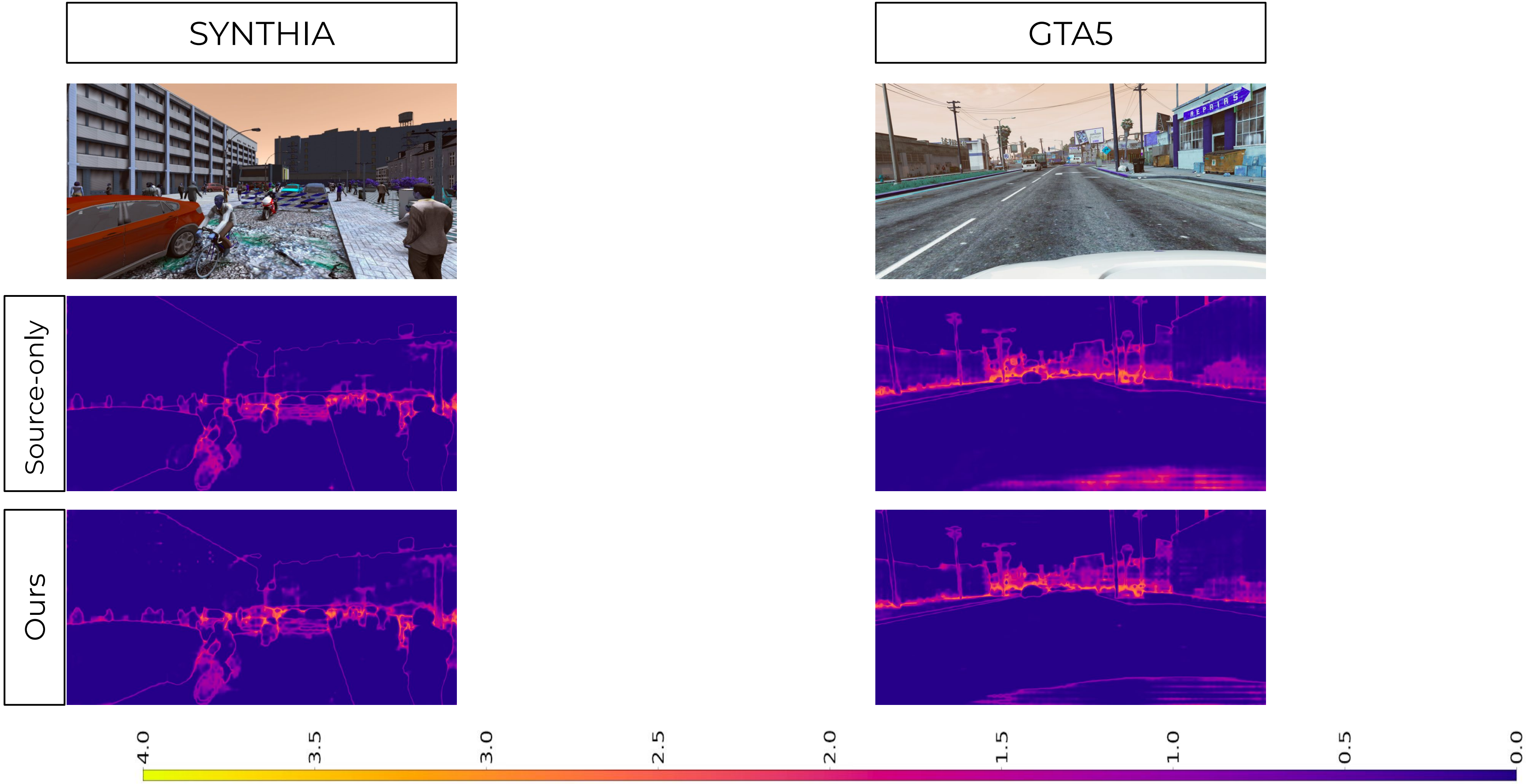
[1] Vu, Tuan-Hung, et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
[2] Zou, Yang, et al. "Confidence regularized self-training." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

SYNTHIA -> Cityscapes

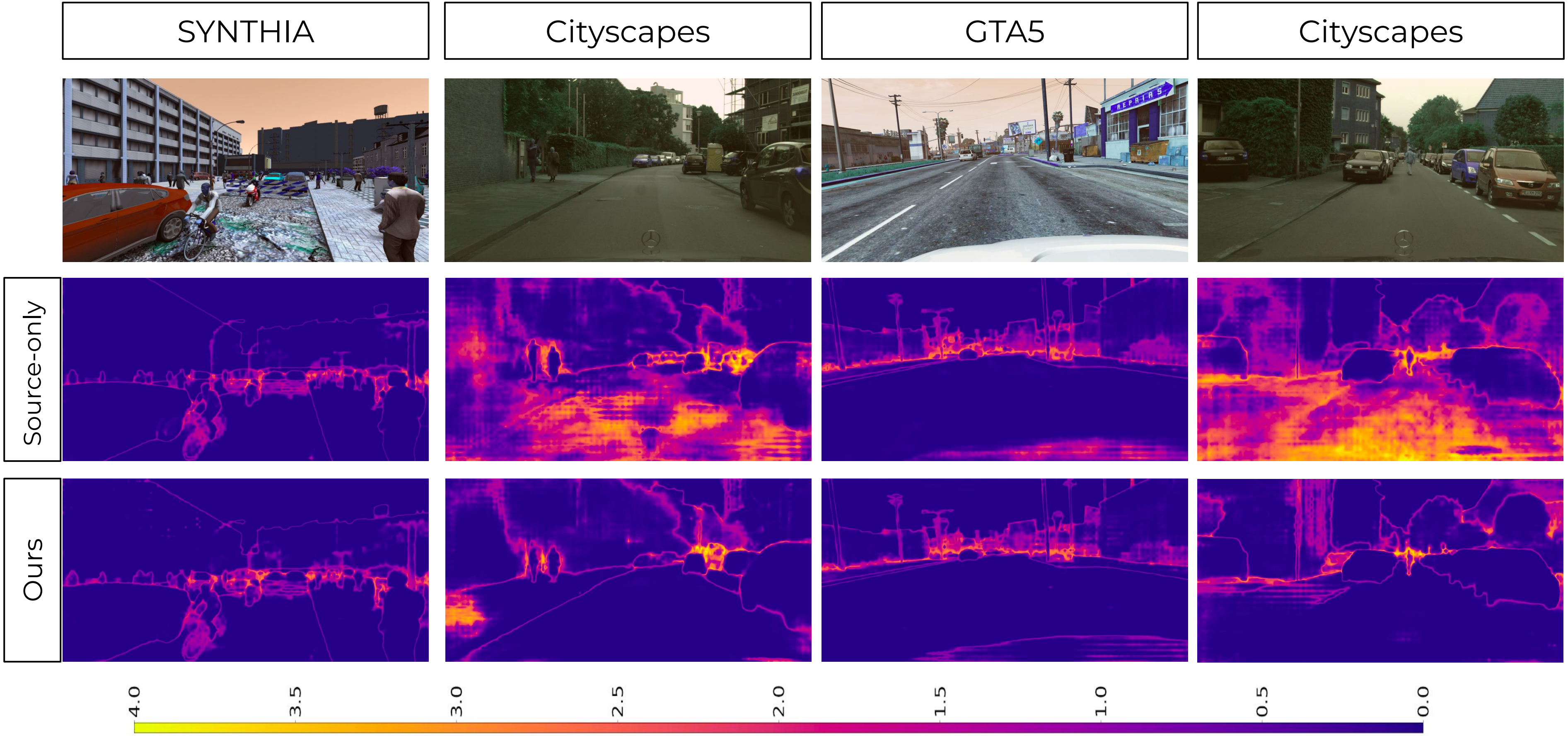


SYNTHIA -> Cityscapes

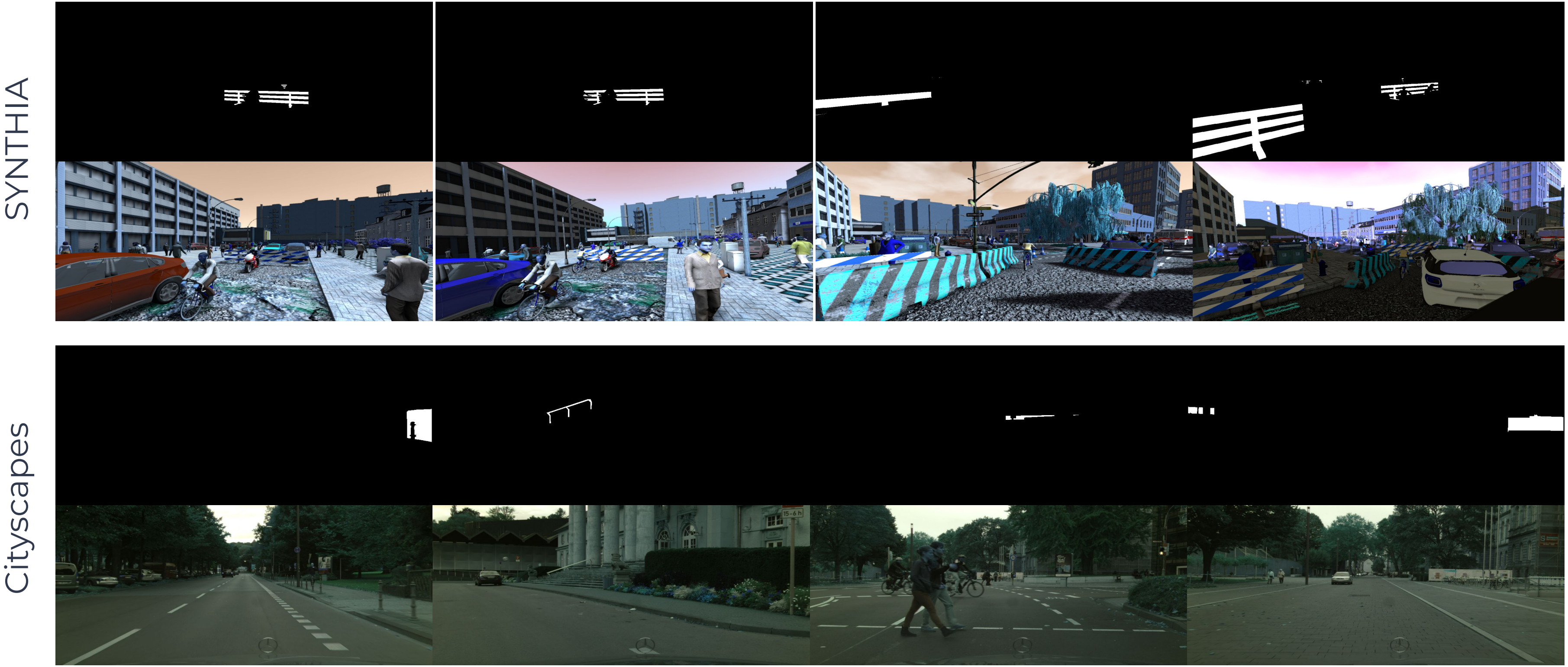
Entropy of Predictions



Entropy of Predictions



Failure Cases



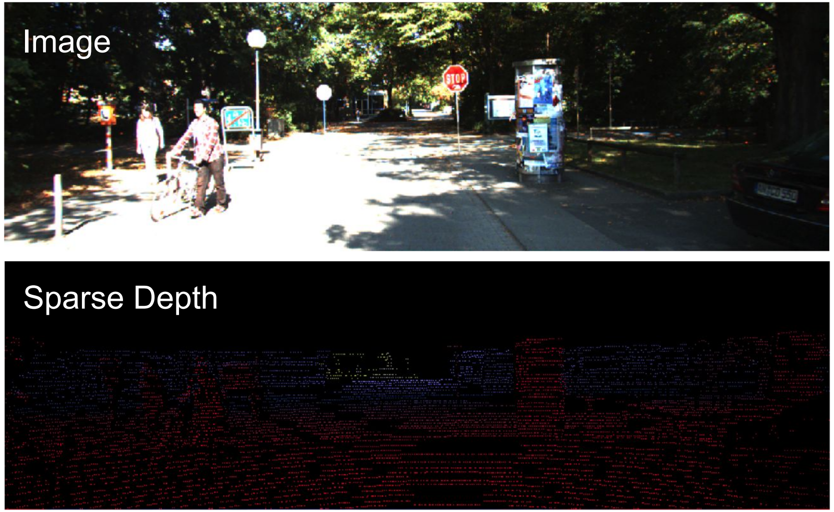
Learning Topology from Synthetic Data for Unsupervised Depth Completion

[1] Alex Wong, Safa Cicek, Stefano Soatto, Learning Topology from Synthetic Data for Unsupervised Depth Completion, IEEE Robotics and Automation Letters (RAL). 2021.

Sparse to Dense Depth Completion

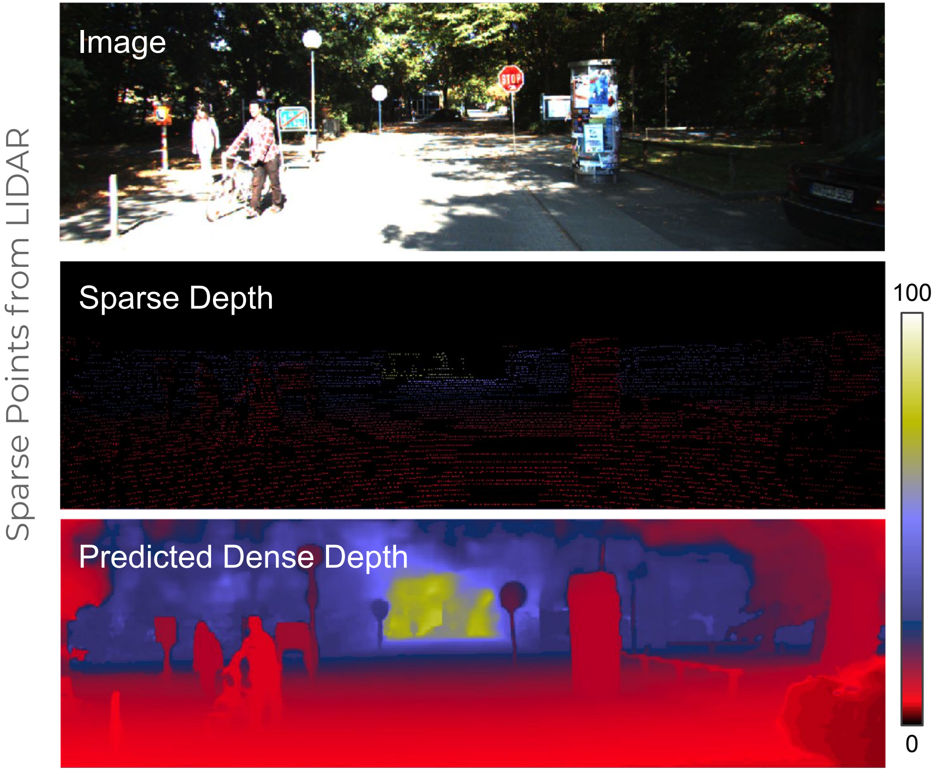
*VIO: Visual Inertial Odometry

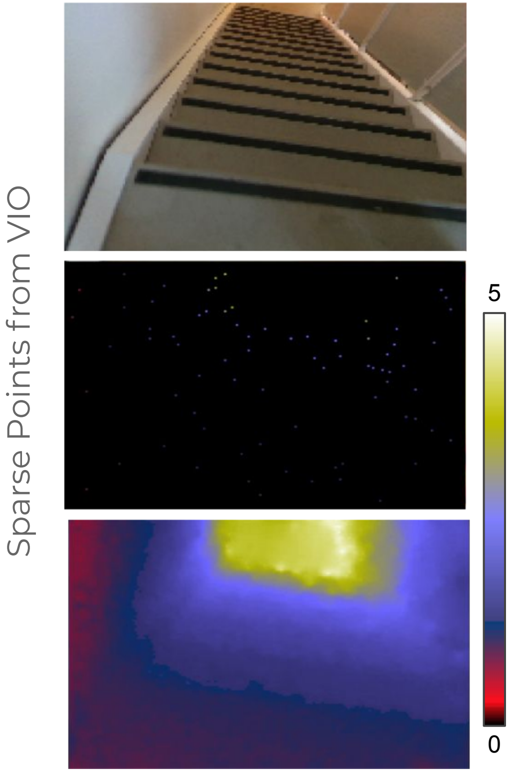
Sparse Points from LIDAR



Sparse Points from VIO*

Sparse to Dense Depth Completion

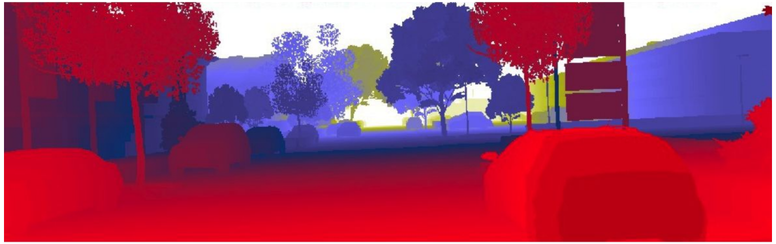




Unsupervised Domain Adaptation (UDA)

Synthetic Source

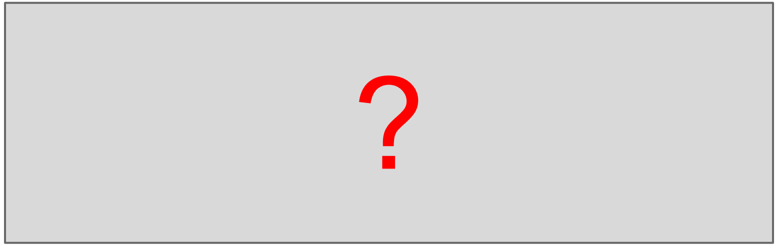
[2]



$$(x^s, y^s) \sim P^s$$

Real Target

[1]



$$x^t \sim P^t_x$$

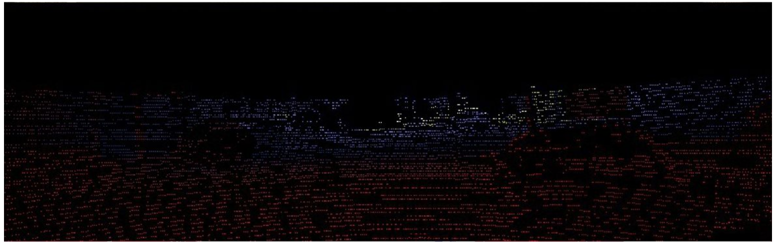
$$KL(P^s||P^t) \gg 0$$

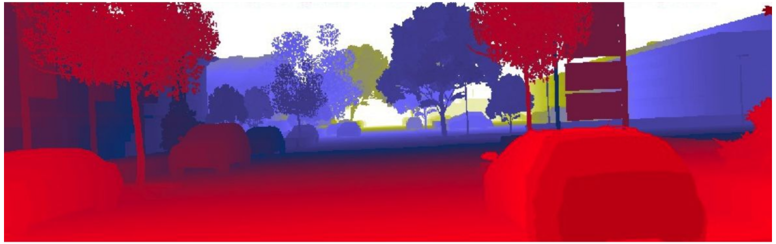
[1] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant cnns. 3DV 2017.
[2] Y. Cabon, N. Murray, M. Humenberger. Virtual KITTI 2. Preprint 2020.

Bypassing the Photometric Domain Gap

Synthetic Source

[2]

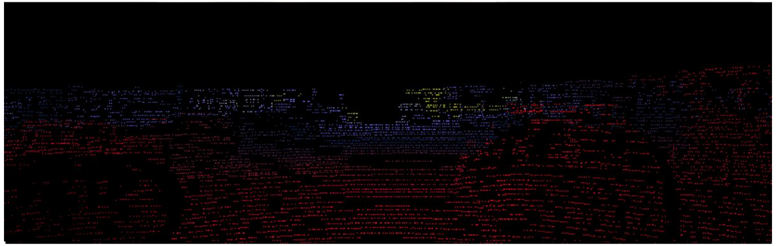


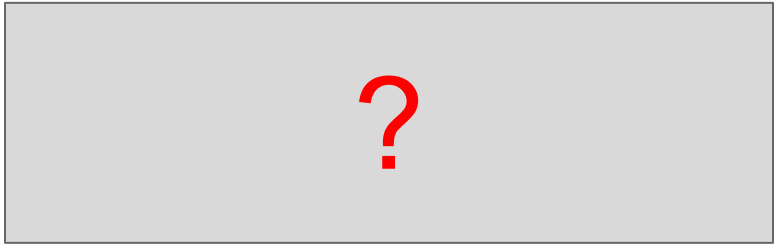


$$(x^s, y^s) \sim P^s$$

Real Target

[1]





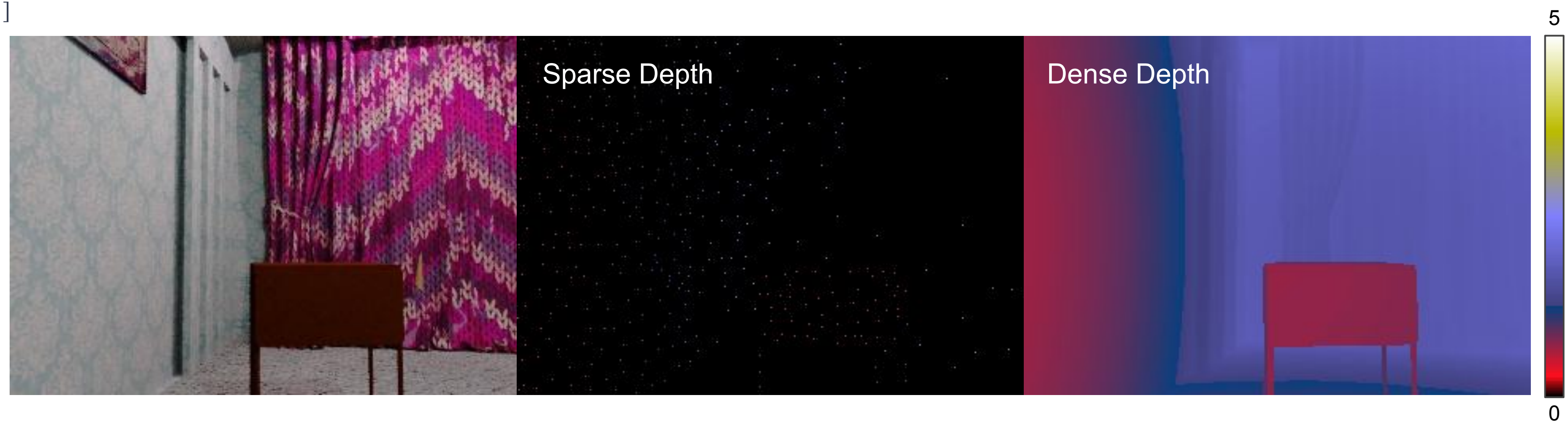
$$x^t \sim P^t_x$$

$$KL(P^s || P^t) \approx 0$$

[1] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant cnns. 3DV 2017.
[2] Y. Cabon, N. Murray, M. Humenberger. Virtual KITTI 2. Preprint 2020.

Bypassing the Photometric Domain Gap

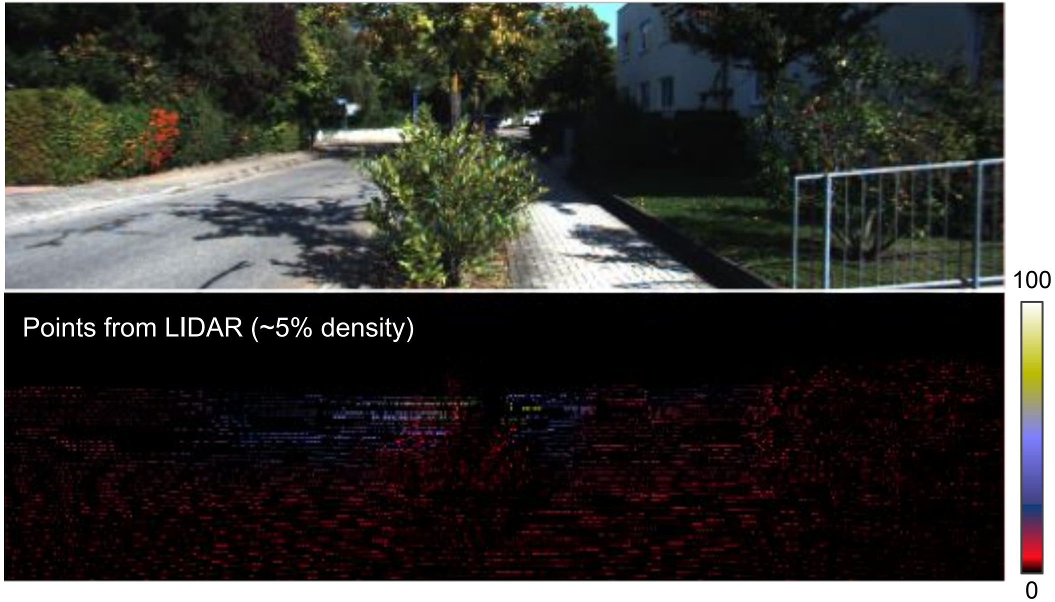
[1]

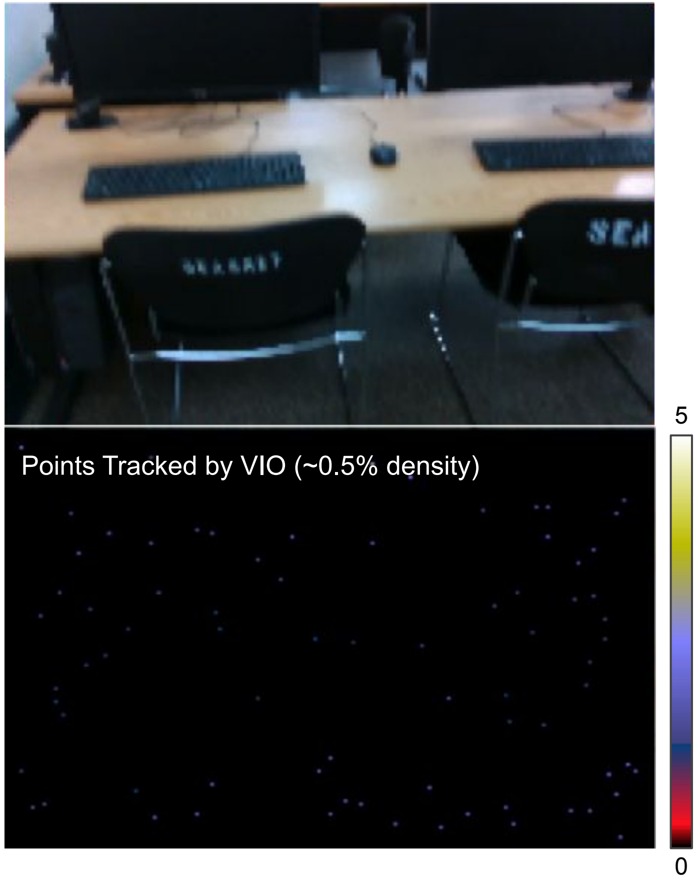


Can we learn to infer the dense topology of the scene given only sparse points?

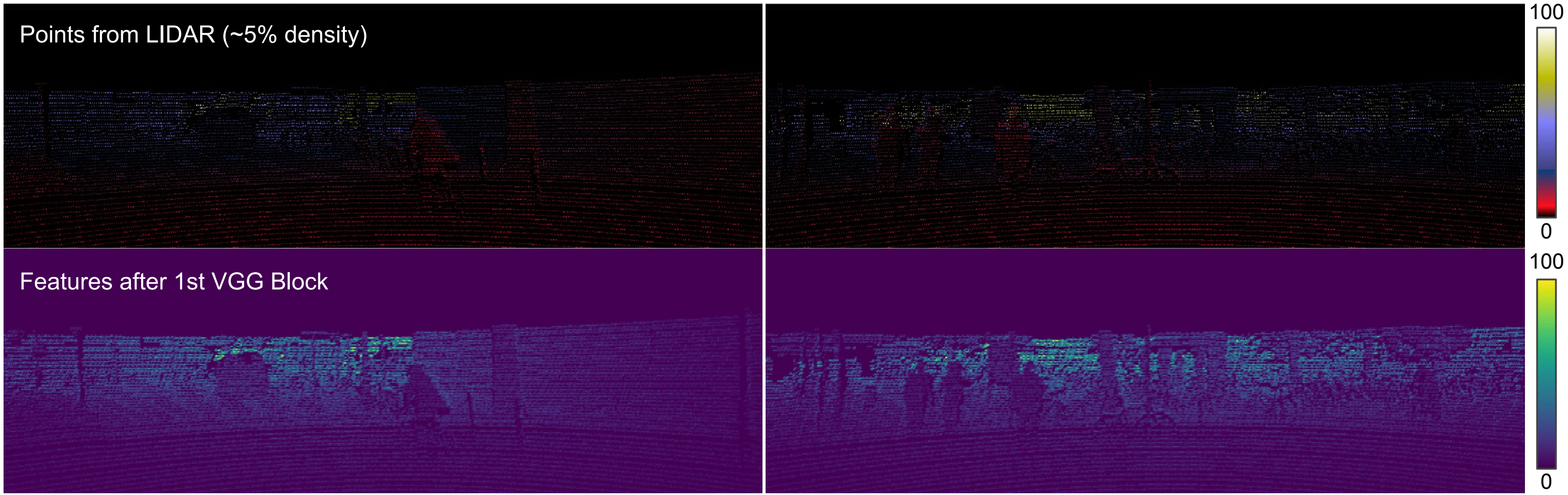
[1] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with groundtruth. Preprint 2016.

The Sparsity Problem



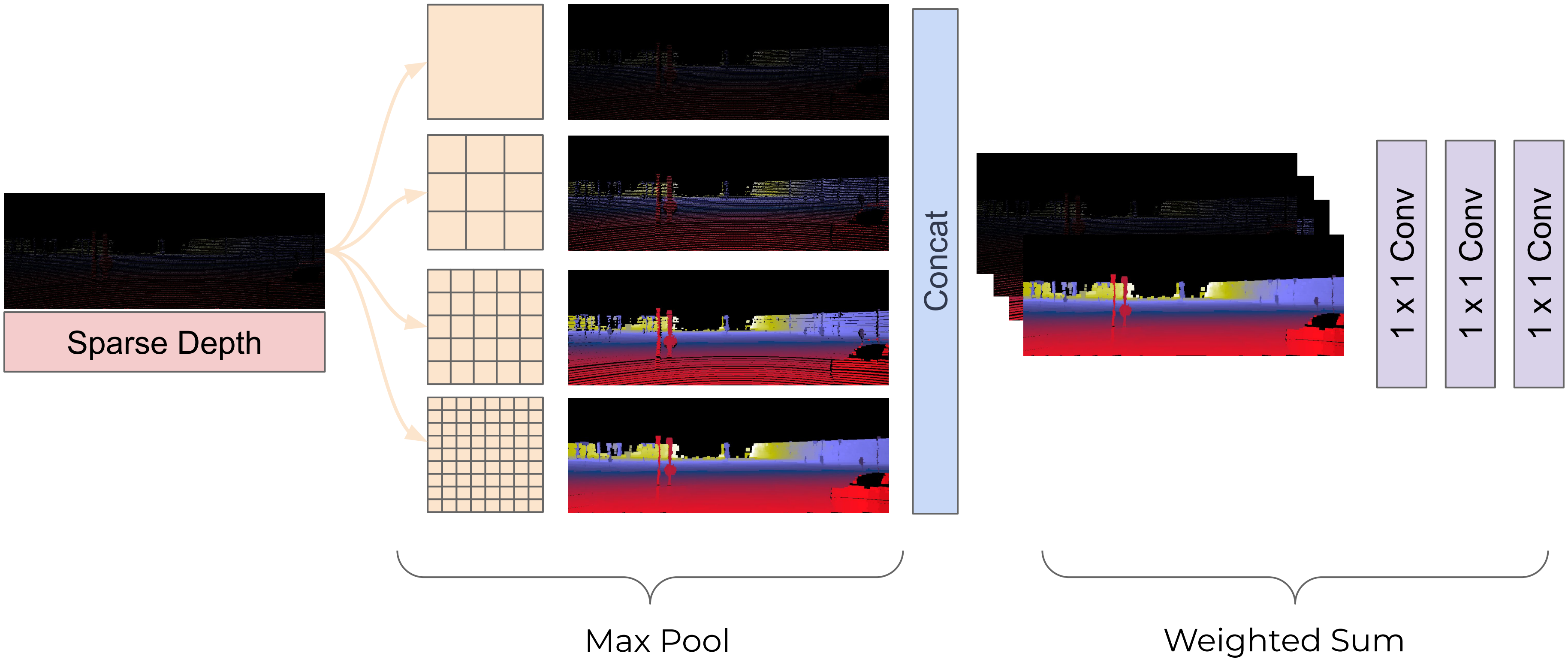


The Sparsity Problem



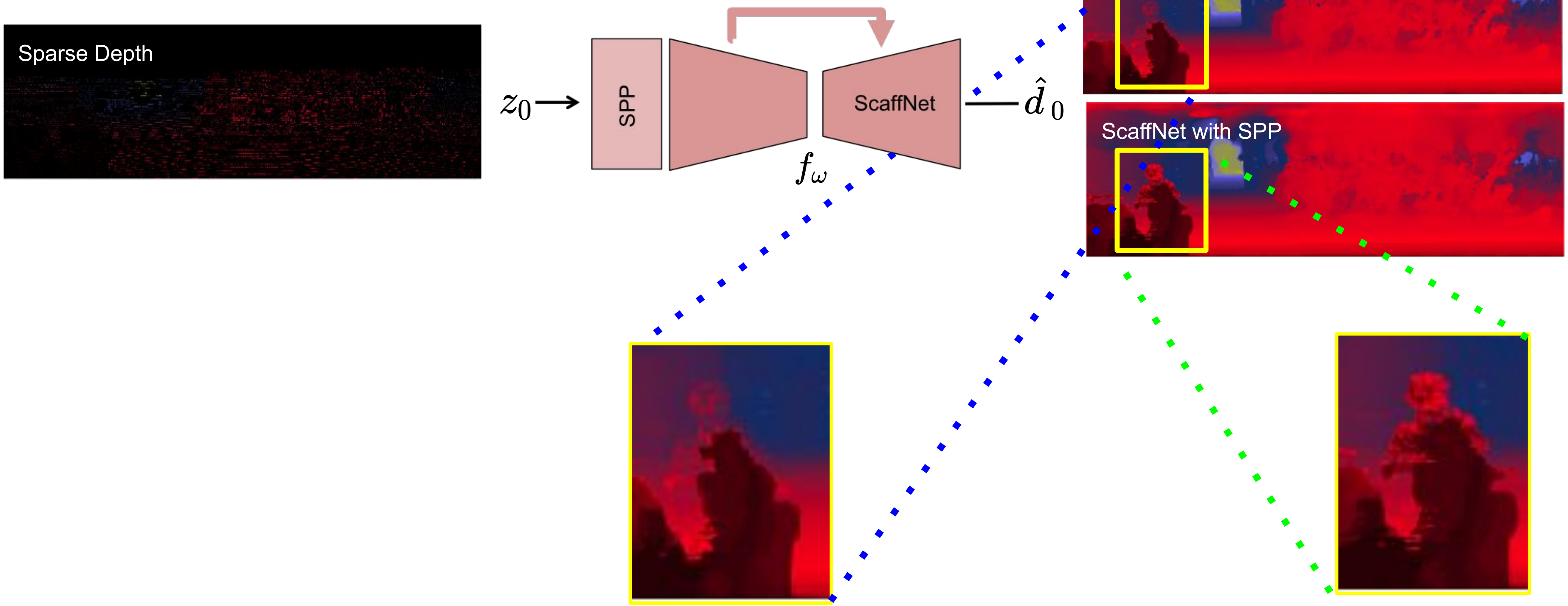
Feature maps are still sparse after the first convolution block.

Spatial Pyramid Pooling (SPP)



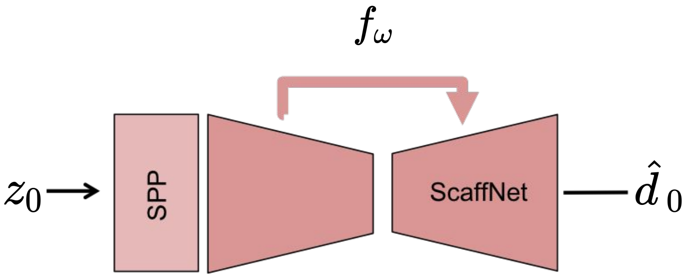
ScaffNet

$$l_0 = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left| \frac{f_\omega(z_0(x)) - d_0(x)}{d_0(x)} \right|$$

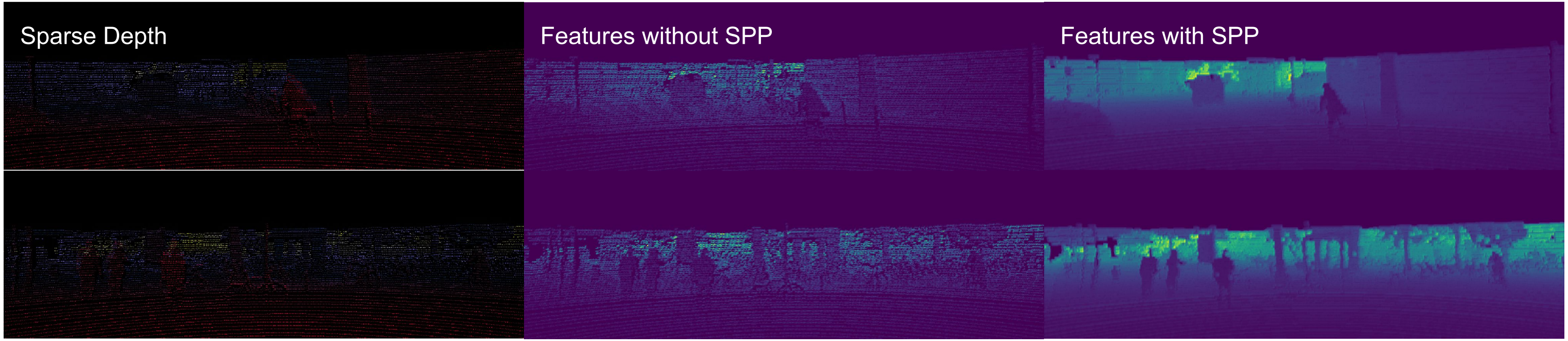


ScaffNet

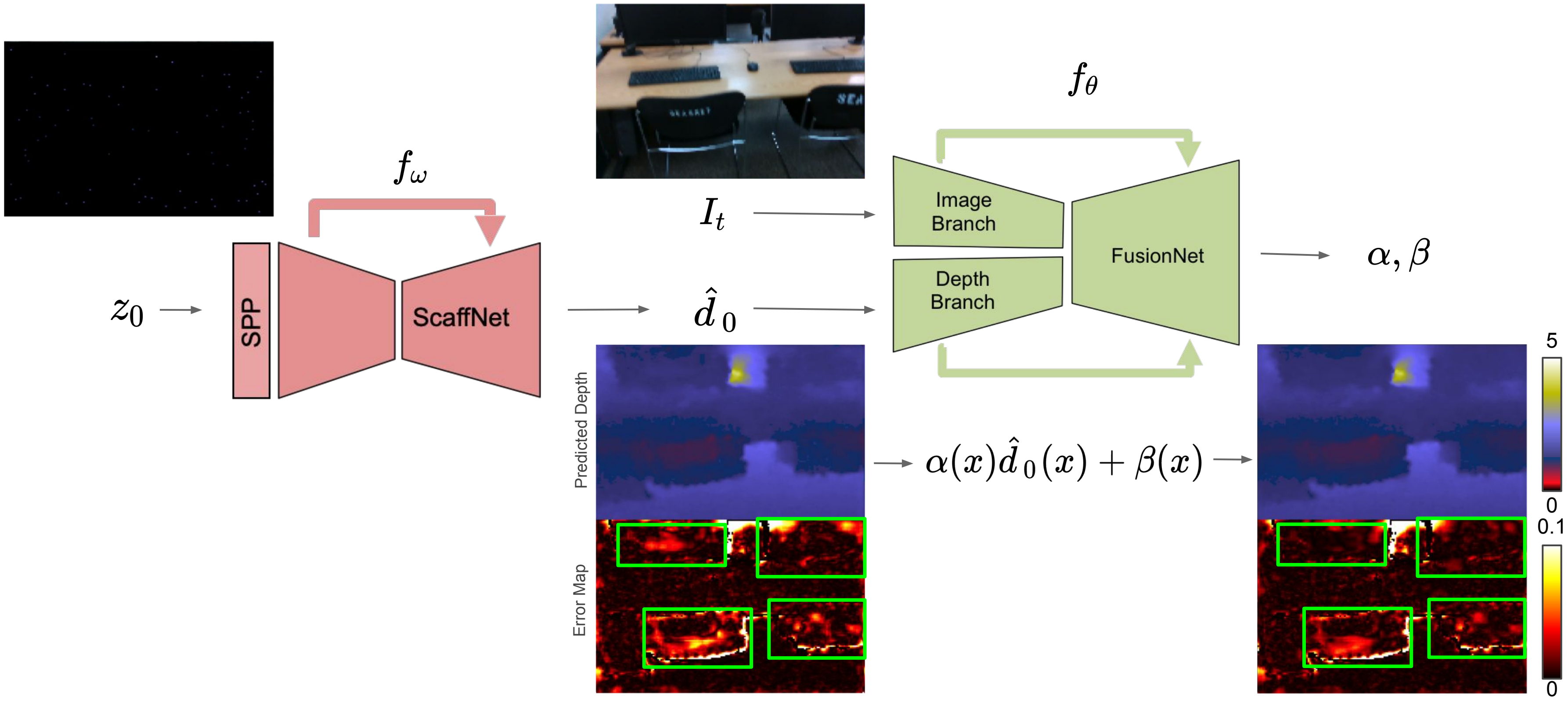

Sparse Depth


f_ω
z_0
SPP
ScaffNet
$\hat{d}_0$

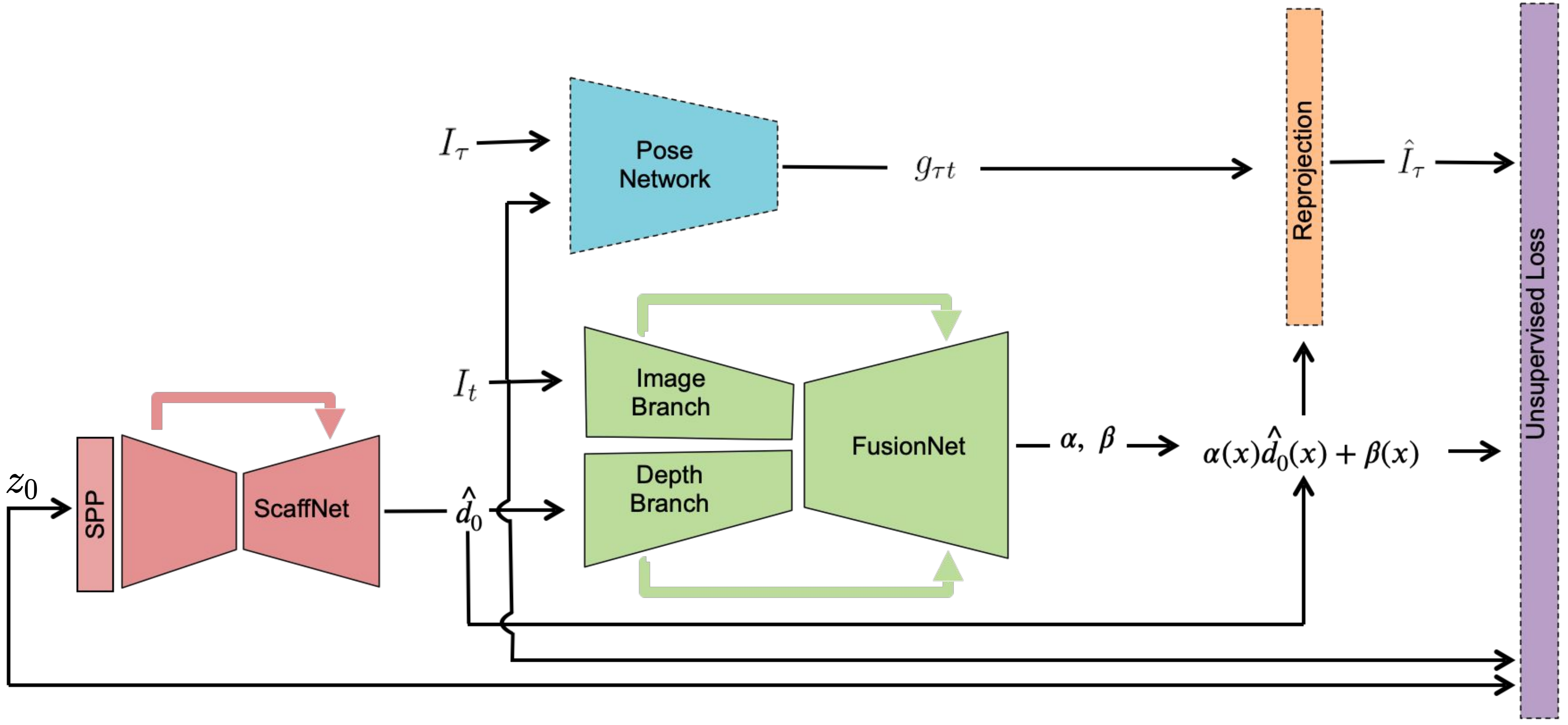

ScaffNet without SPP
ScaffNet with SPP


Sparse Depth
Features without SPP
Features with SPP

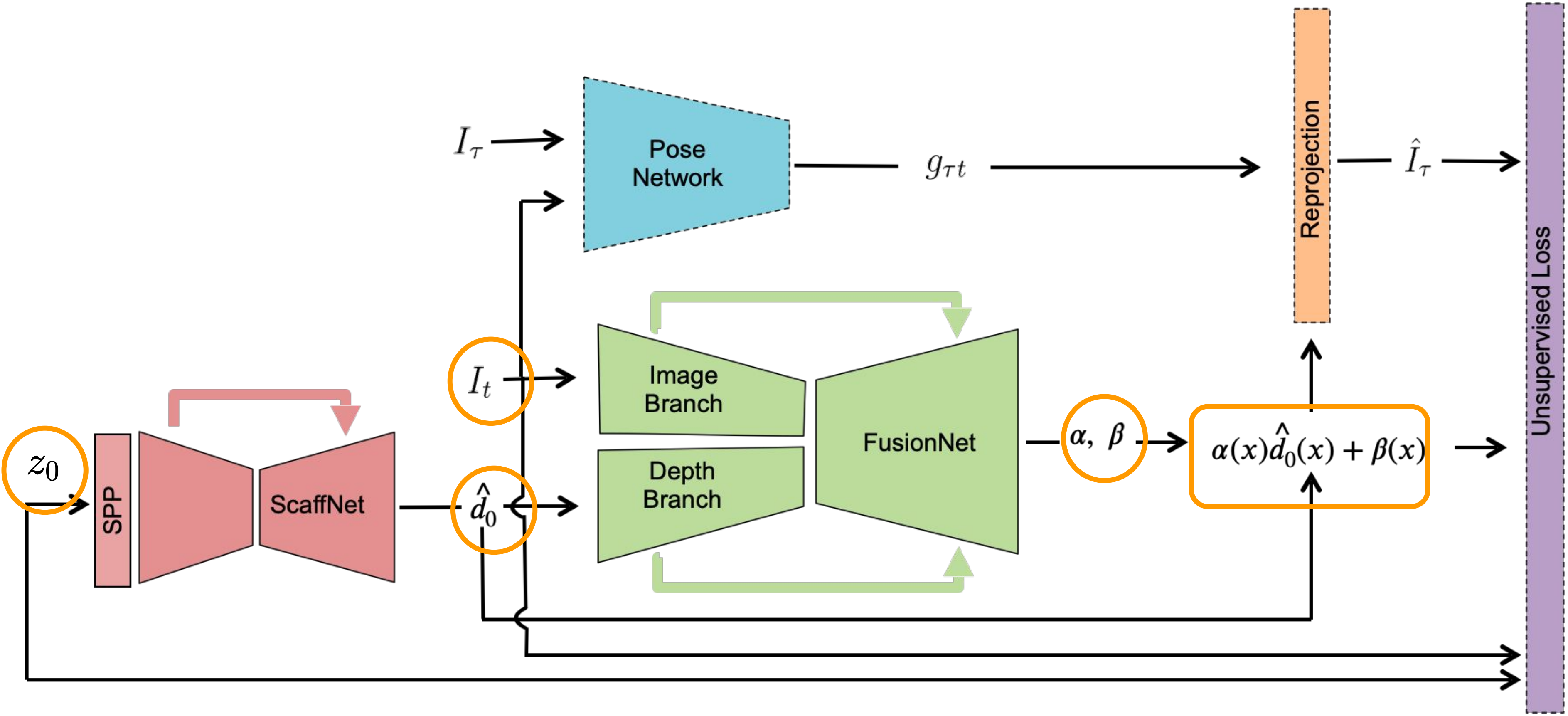
Bringing the Image Back



FusionNet


I_τ
Pose Network
$g_{\tau t}$
Reprojection
$\hat{I}_\tau$
Unsupervised Loss
I_t
Image Branch
Depth Branch
FusionNet
$\alpha,\ \beta$
$\alpha(x)\hat{d}_0(x) + \beta(x)$
z_0
SPP
ScaffNet
$\hat{d}_0$

FusionNet



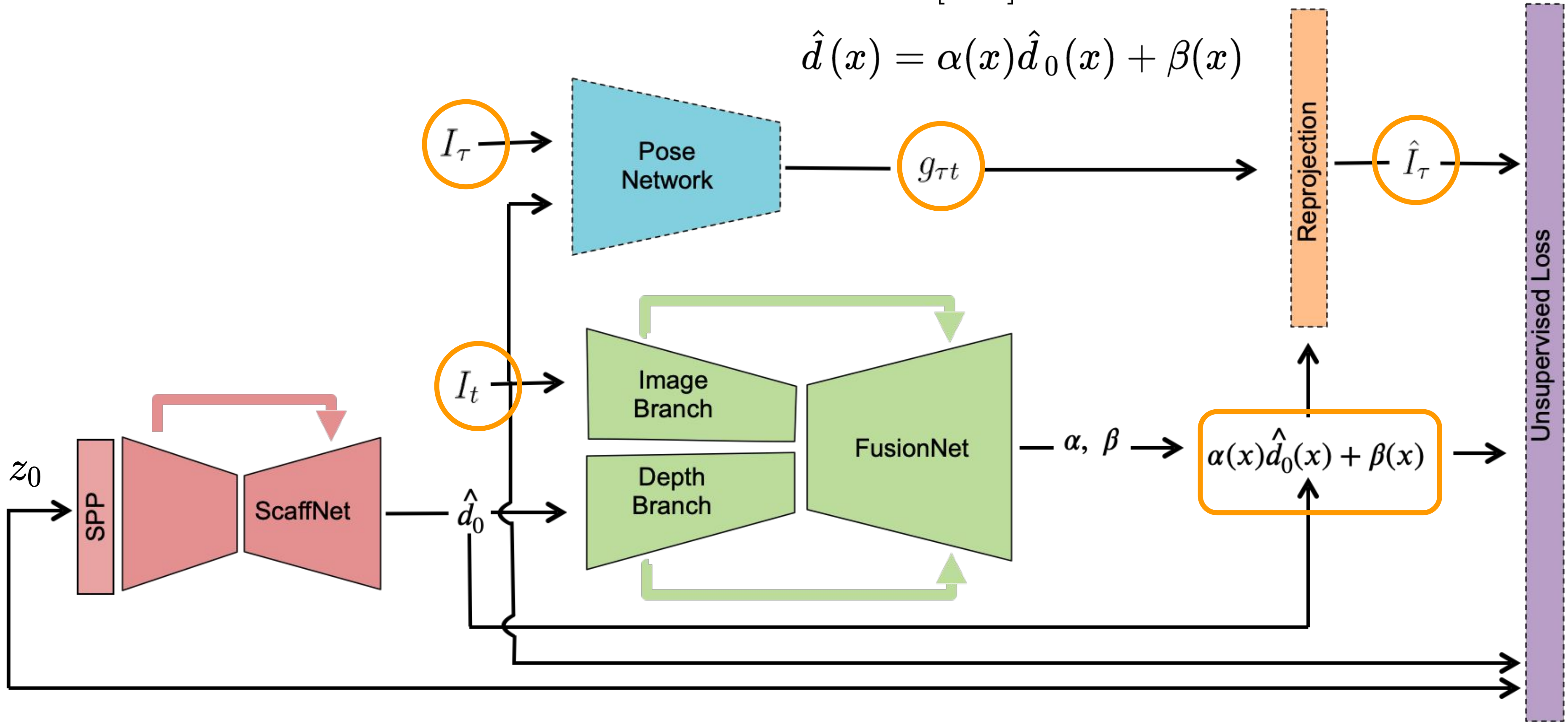
I_τ
Pose Network
$g_{\tau t}$
Reprojection
$\hat{I}_\tau$
Unsupervised Loss
I_t
Image Branch
Depth Branch
FusionNet
$\alpha,\ \beta$
$\alpha(x)\hat{d}_0(x) + \beta(x)$
z_0
SPP
ScaffNet
$\hat{d}_0$

FusionNet

$$\hat{I}_\tau(x) = I_\tau(\pi g_{\tau t} K^{-1} \bar{x} \hat{d}(x))$$

$$\bar{x} = [x \;\; 1]^\top$$

$$\hat{d}(x) = \alpha(x)\hat{d}_0(x) + \beta(x)$$

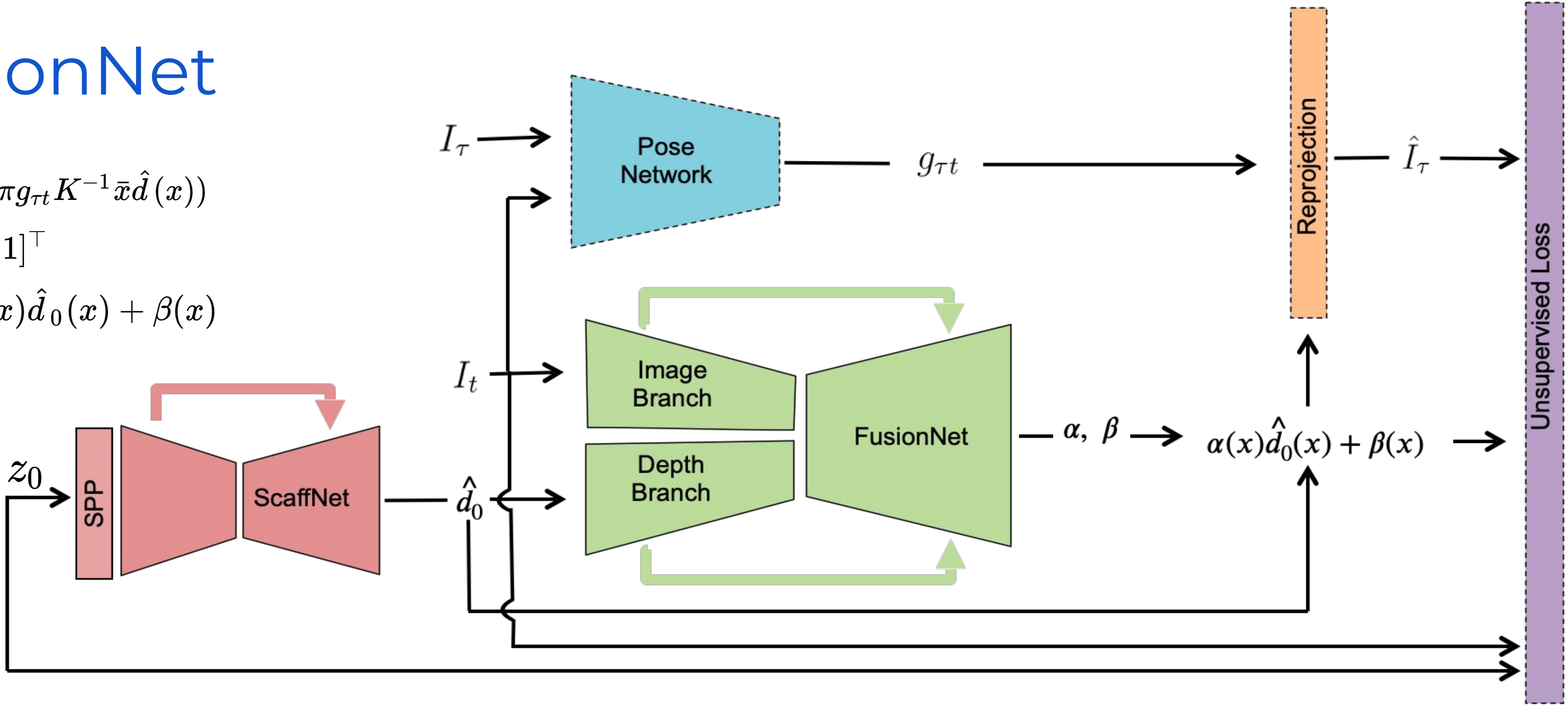


FusionNet

$$\hat{I}_\tau(x) = I_\tau(\pi g_{\tau t} K^{-1} \bar{x} \hat{d}(x))$$

$$\bar{x} = [x \;\; 1]^\top$$

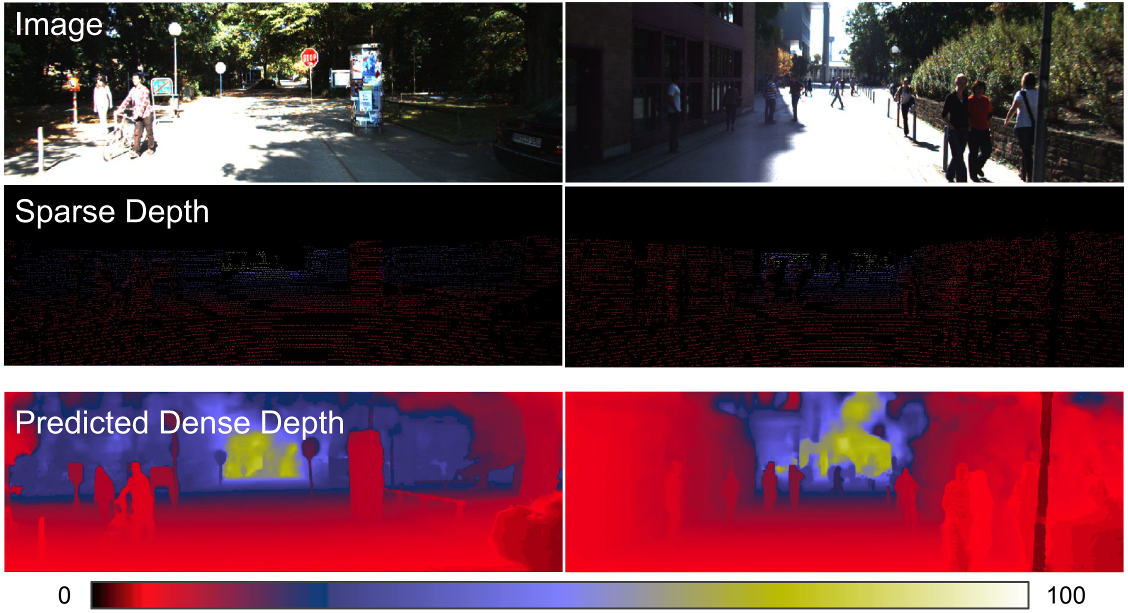
$$\hat{d}(x) = \alpha(x)\hat{d}_0(x) + \beta(x)$$



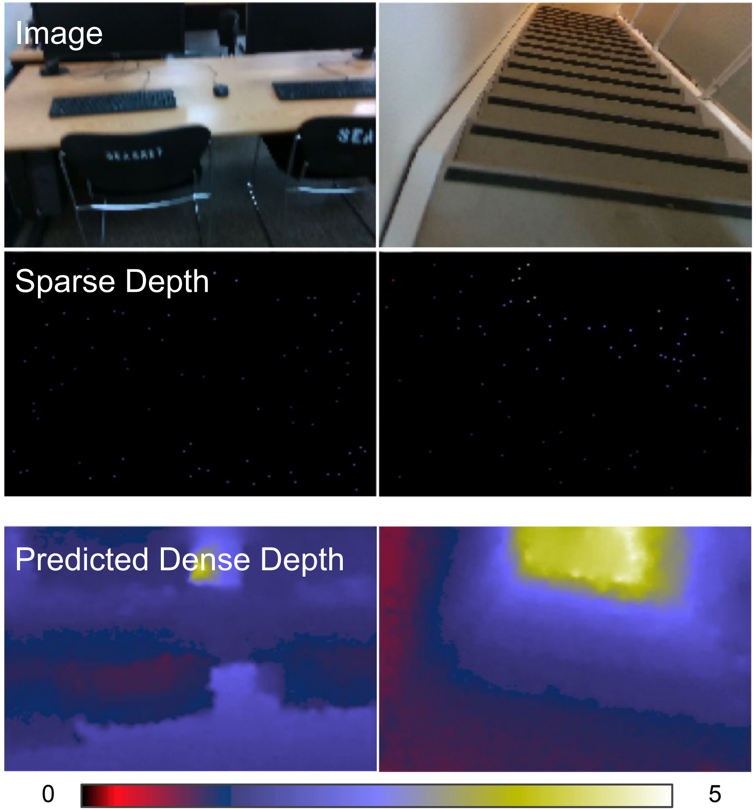
$$\mathcal{L} = \underbrace{w_{ph} \frac{1}{|\Omega|} \ell(I_t(x), \hat{I}_\tau(x))}_{\text{photometric consistency}} + \underbrace{w_{sz} \frac{1}{|\Omega_z|} |z_0(x) - \hat{d}(x)|}_{\text{sparse depth consistency}} + \underbrace{w_{sm} \frac{1}{|\Omega|} \lambda |\nabla \hat{d}(x)|}_{\text{local smoothness}} + \underbrace{w_{tp} \frac{1}{\sum_{x \in \Omega} W(x)} \sum_{x \in \Omega} W(x) |\hat{d}(x) - \hat{d}_0(x)|}_{\text{topology prior}}$$

Qualitative Results

[1] KITTI



[2] VOID



[1] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant cnns. 3DV 2017.
[2] X.Fei, A. Wong, S. Soatto. Geo-Supervised Depth Prediction. R-AL 2019 and ICRA 2019.

Quantitative Results

Method	Parameters	MAE	RMSE	iMAE	iRMSE
ScaffNet	**~1.4M**	**318.41**	**1425.53**	**1.39**	**5.01**
[1]	~27.8M	358.92	1384.85	1.60	4.32
[2]	~18.8M	347.17	1310.03	n/a	n/a
[3]	~9.7M	305.06	1239.06	1.21	3.71
FusionNet	**~7.8M**	**286.35**	**1182.81**	**1.18**	**3.55**

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert^2\right)^{1/2}$
iMAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert$
iRMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert^2\right)^{1/2}$

[1] F. Ma, G. V. Cavalheiro, S. Karaman. Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from LiDAR and Monocular Camera. ICRA 2019.
[2] Y. Yang, A. Wong, S. Soatto. Dense Depth Posterior (DDP) from Single Image and Sparse Range. CVPR 2019.
[3] A. Wong. X. Fei, S. Tsuei, S. Soatto. Unsupervised Depth Completion from Visual Inertial Odometry. R-AL 2020, and ICRA, 2020.

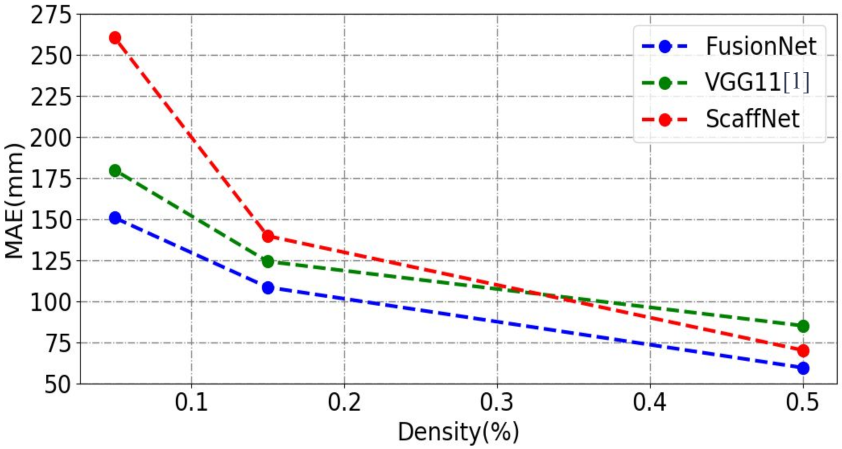
Quantitative Results -- Indoor

Method	Parameters	MAE	RMSE	iMAE	iRMSE
[1]	~27.8M	198.76	260.67	88.07	114.96
[2]	~18.8M	151.86	222.36	74.59	112.36
[3]	~9.7M	85.05	169.79	48.92	104.02
ScaffNet	**~1.4M**	**70.16**	**156.99**	**42.78**	**91.48**
FusionNet	**~7.8M**	**59.53**	**119.14**	**35.72**	**68.36**

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert^2\right)^{1/2}$
iMAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert$
iRMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert^2\right)^{1/2}$

[1] F. Ma, G. V. Cavalheiro, S. Karaman. Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from LiDAR and Monocular Camera. ICRA 2019.
[2] Y. Yang, A. Wong, S. Soatto. Dense Depth Posterior (DDP) from Single Image and Sparse Range. CVPR 2019.
[3] A. Wong. X. Fei, S. Tsuei, S. Soatto. Unsupervised Depth Completion from Visual Inertial Odometry. R-AL 2020, and ICRA, 2020.

Quantitative Results -- Indoor



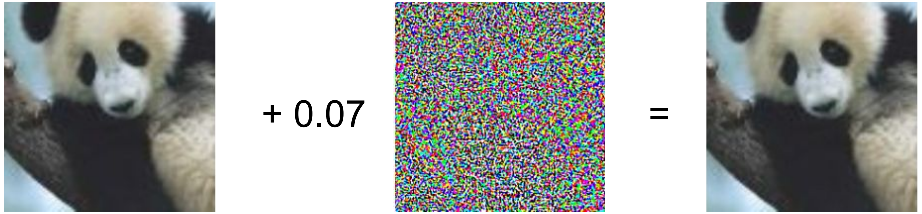
- MAE for various density levels.

[1] A. Wong. X. Fei, S. Tsuei, S. Soatto. Unsupervised Depth Completion from Visual Inertial Odometry. R-AL 2020, and ICRA, 2020.

Targeted Adversarial Perturbations for Monocular Depth Prediction

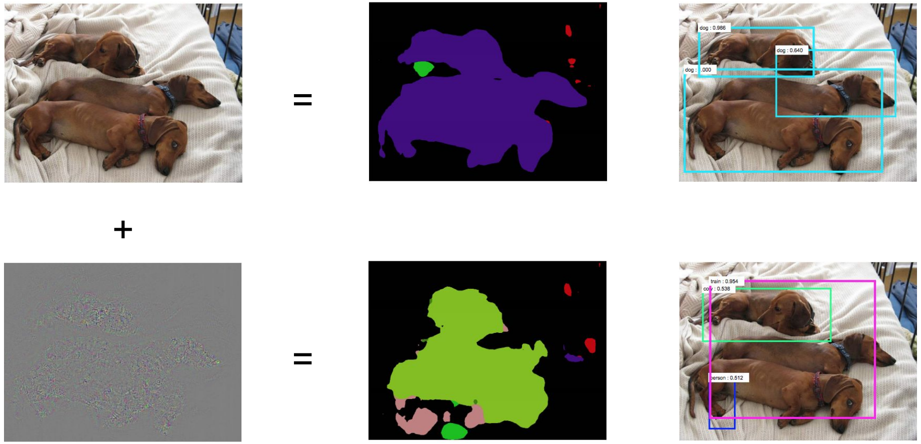
[1] Wong Alex, Safa Cicek, Stefano Soatto, Targeted Adversarial Perturbations for Monocular Depth Prediction. Conference on Neural Information Processing Systems (NeurIPS). 2020.

Adversarial Perturbations



panda gibbon

[1]

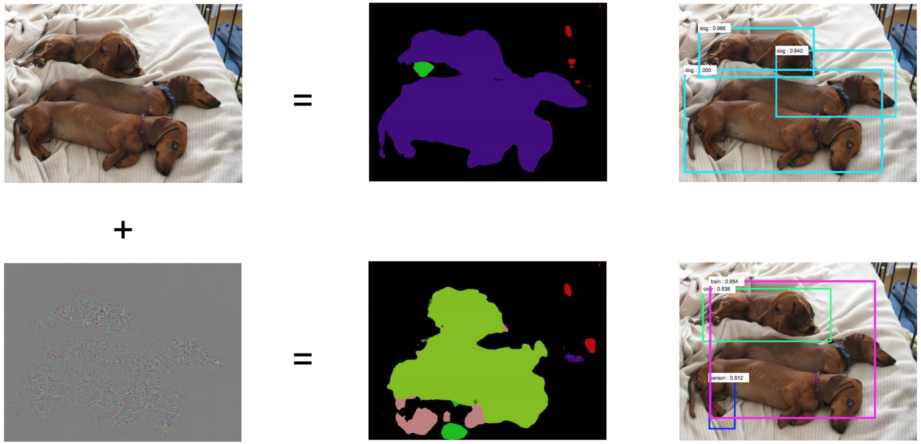


[3]

[1] I. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015.
[2] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. ICCV 2017.

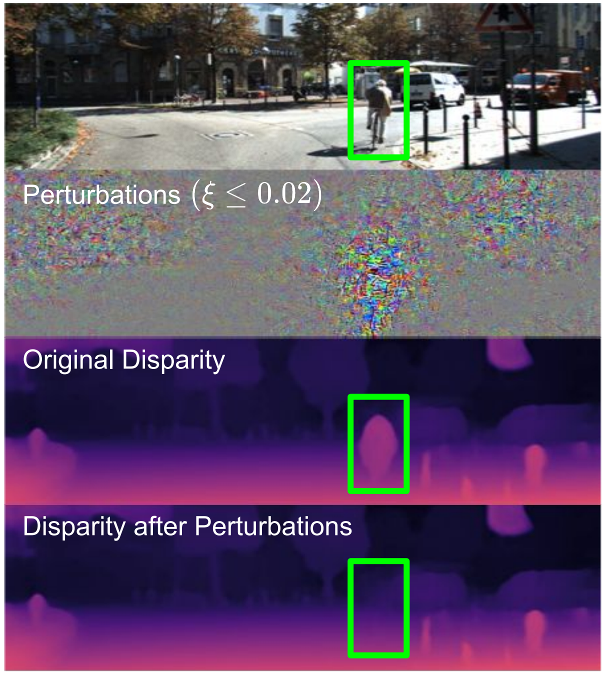
Adversarial Perturbations

[1]



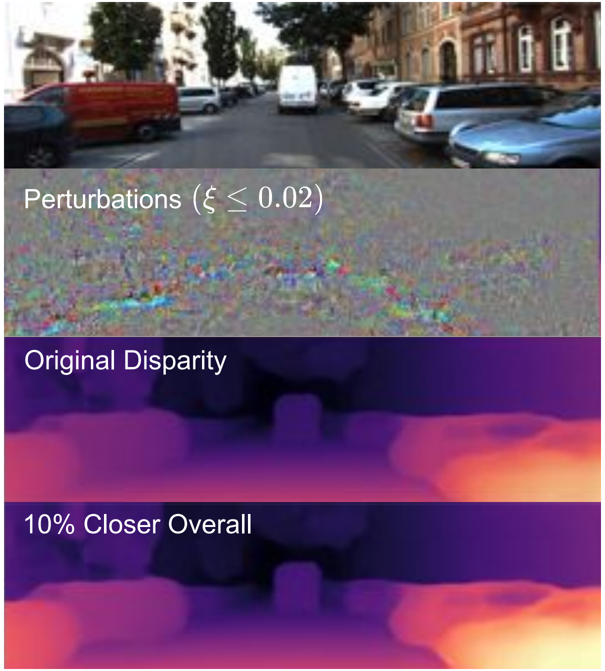
[3]

Targeted Attacks on
Monocular Depth Prediction Networks



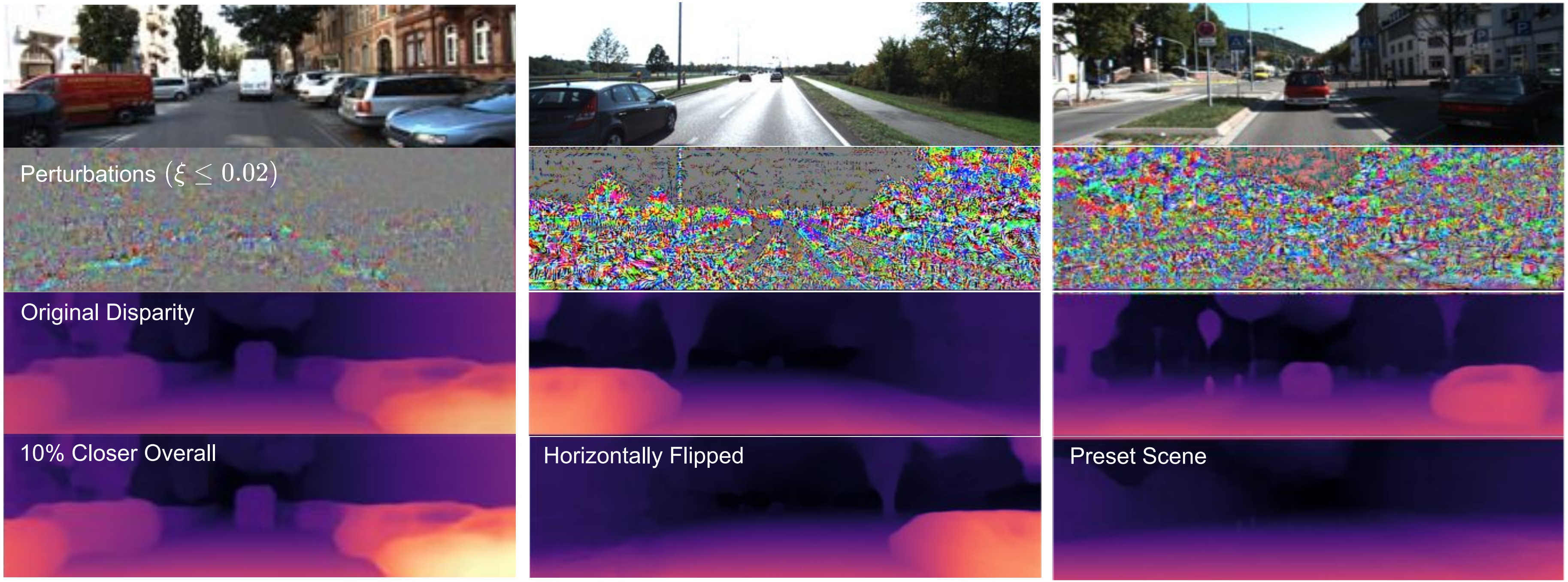
[1] I. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015.
[2] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. ICCV 2017.

Attacking the Entire Scene



(i) scaling the entire scene by a factor of $1 + \alpha$

Attacking the Entire Scene

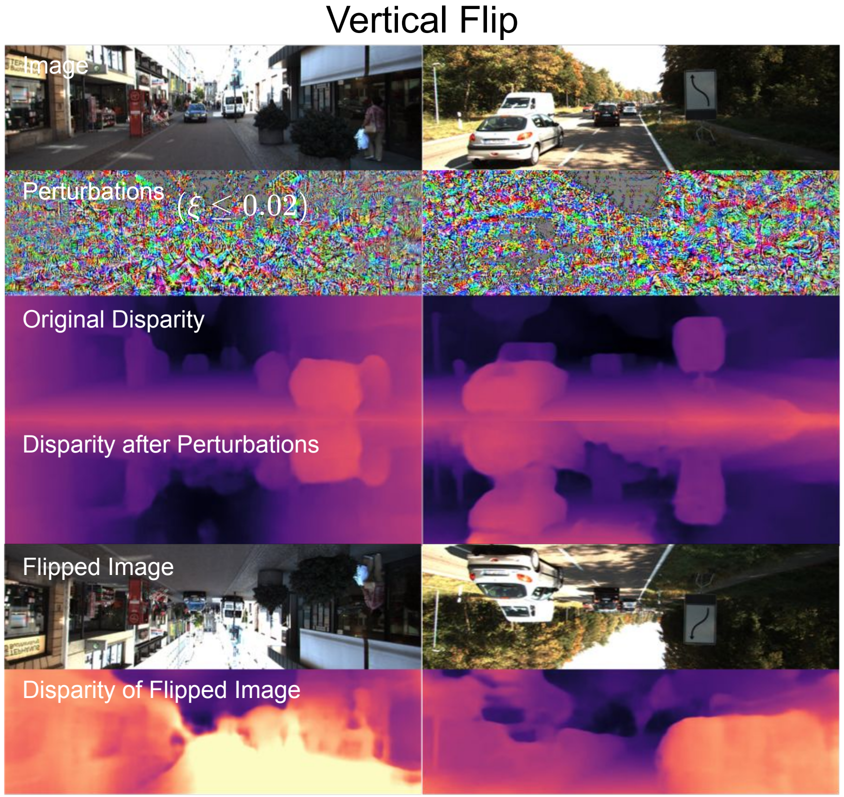


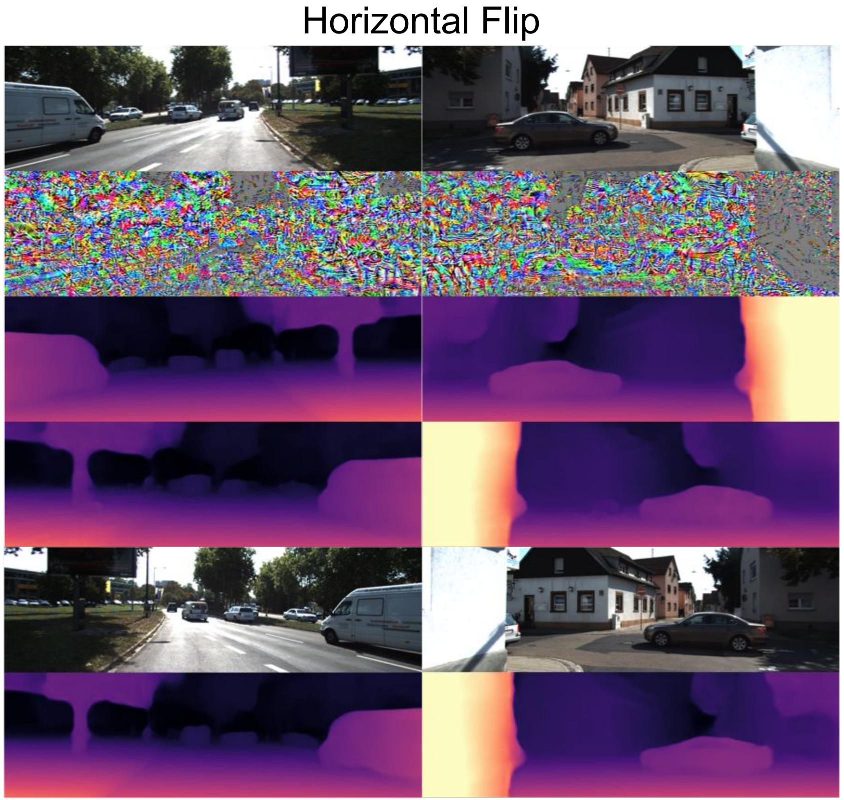
(i) scaling the entire scene by a factor of $1+\alpha$

(ii) symmetrically flipping the entire scene

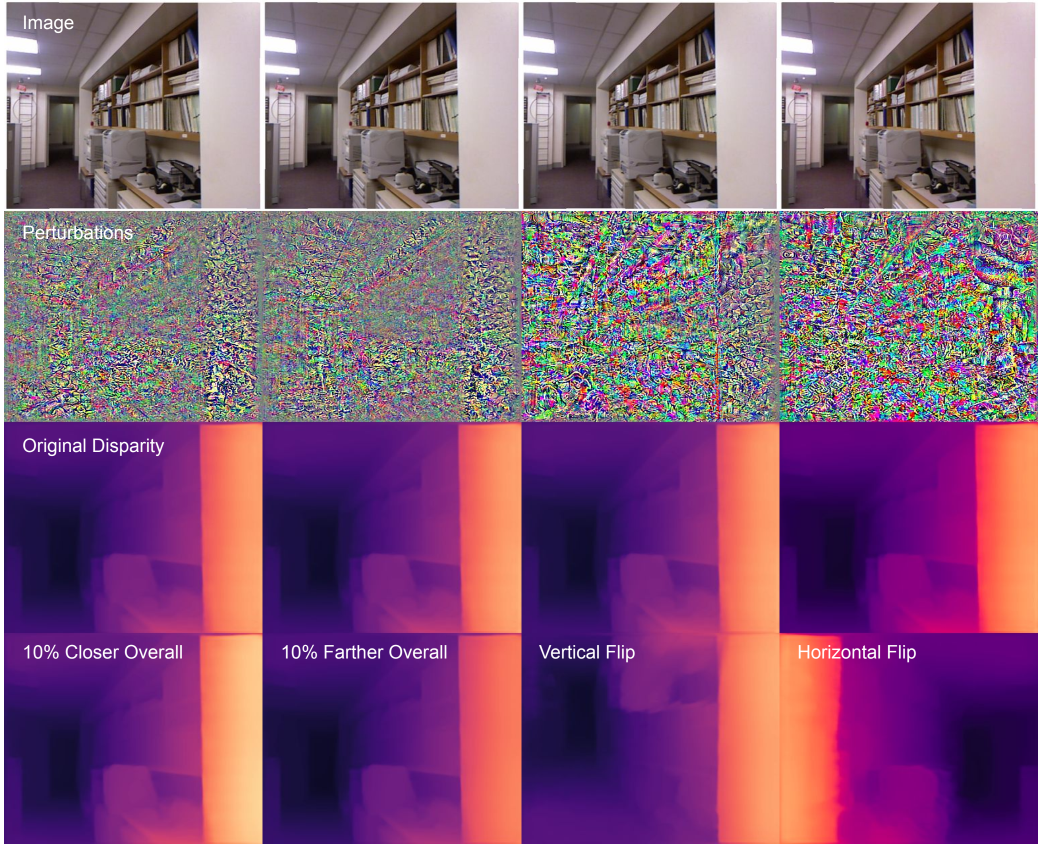
(iii) altering the entire scene to a preset scene

Strong Bias on Scene Orientation





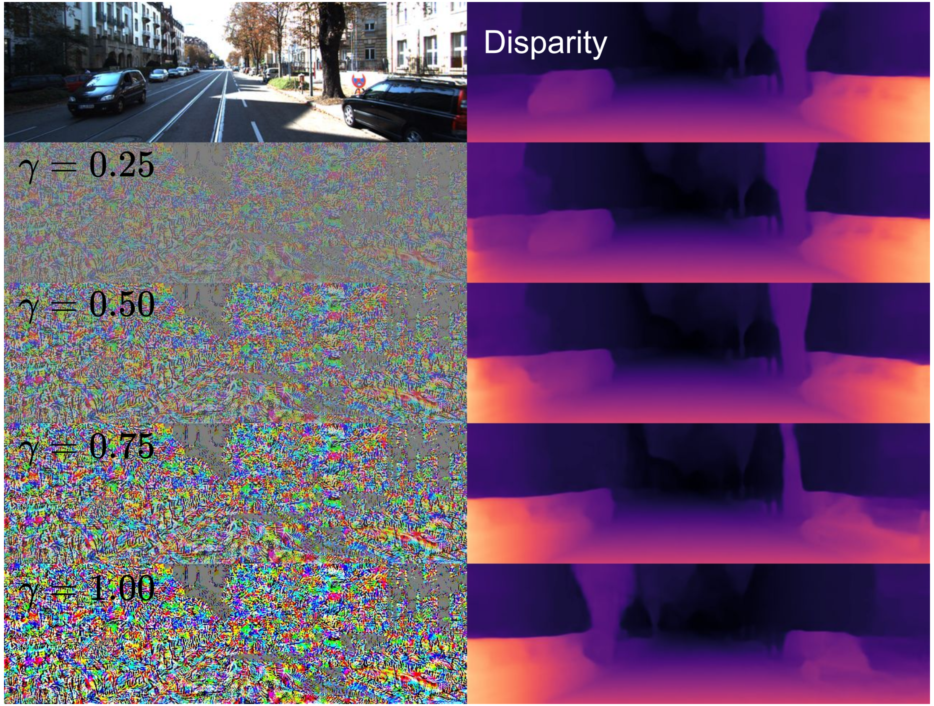
Adversarial Attacks in Indoor Scenes



[1] W. Yin, Y. Liu, C. Shen, Y. Yan. Enforcing geometric constraints of virtual normal for depth prediction. ICCV 2019.

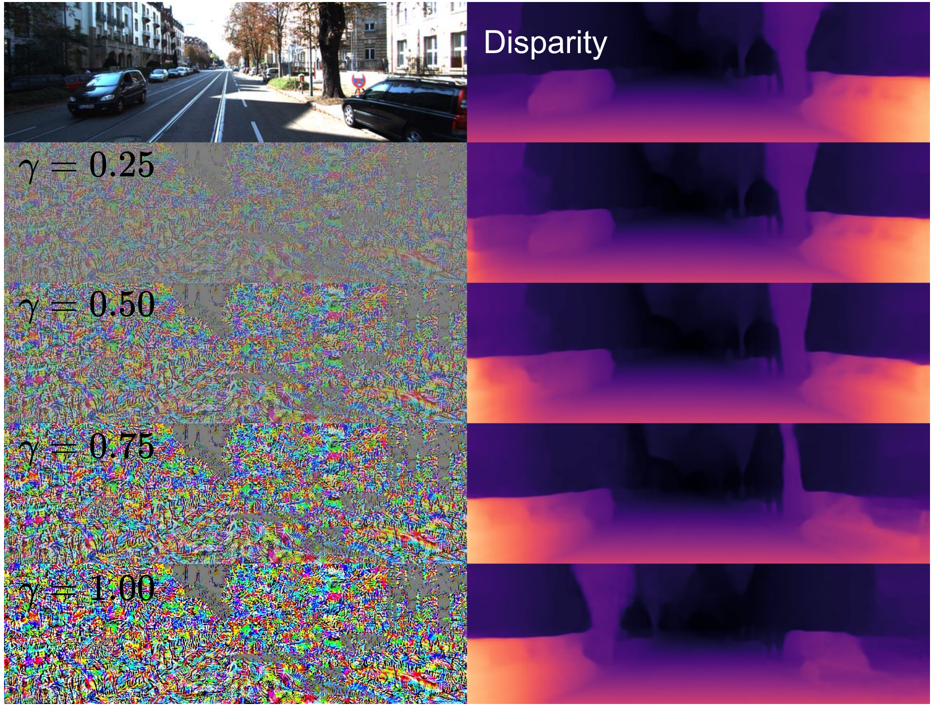
Linear Operations:

$$f_d(x + \gamma\, v(x))$$

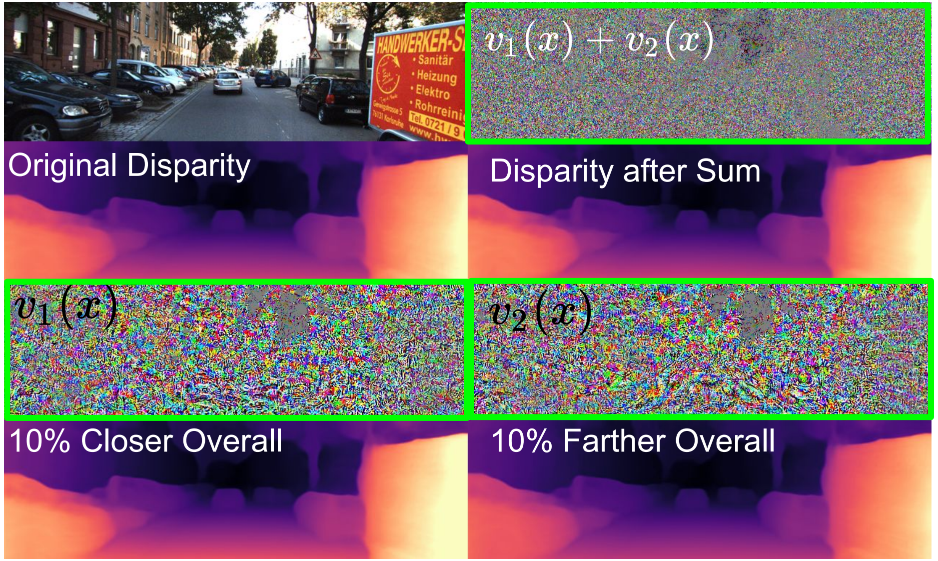


Linear Operations:

$$f_d(x + \gamma\, v(x))$$



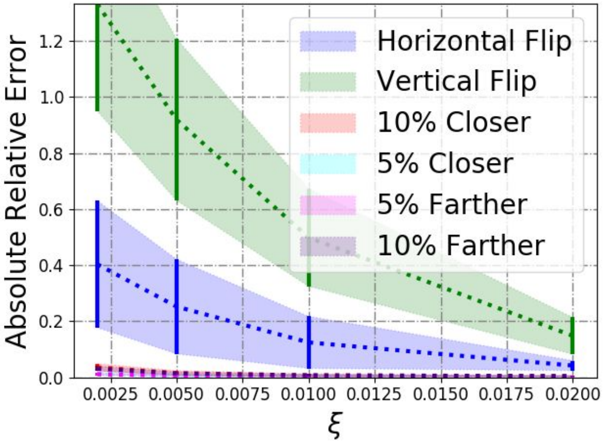
$$f_d(x + v_1(x) + v_2(x))$$



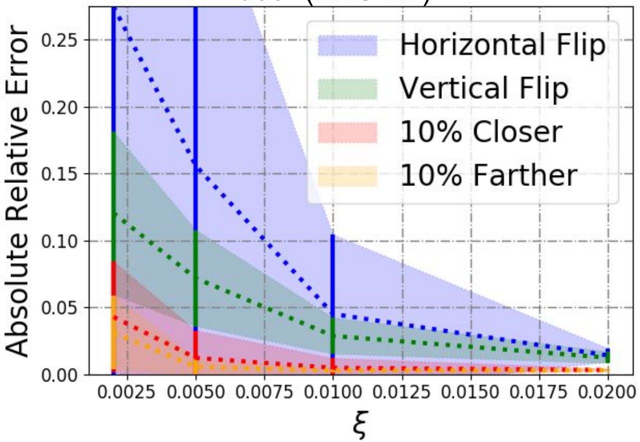
$$||v_1(x)|| \approx ||v_2(x)|| \gg ||v_1(x) + v_2(x)||$$

Quantitative Results
Symmetrically Flipping the Scene

Outdoor Driving (KITTI)

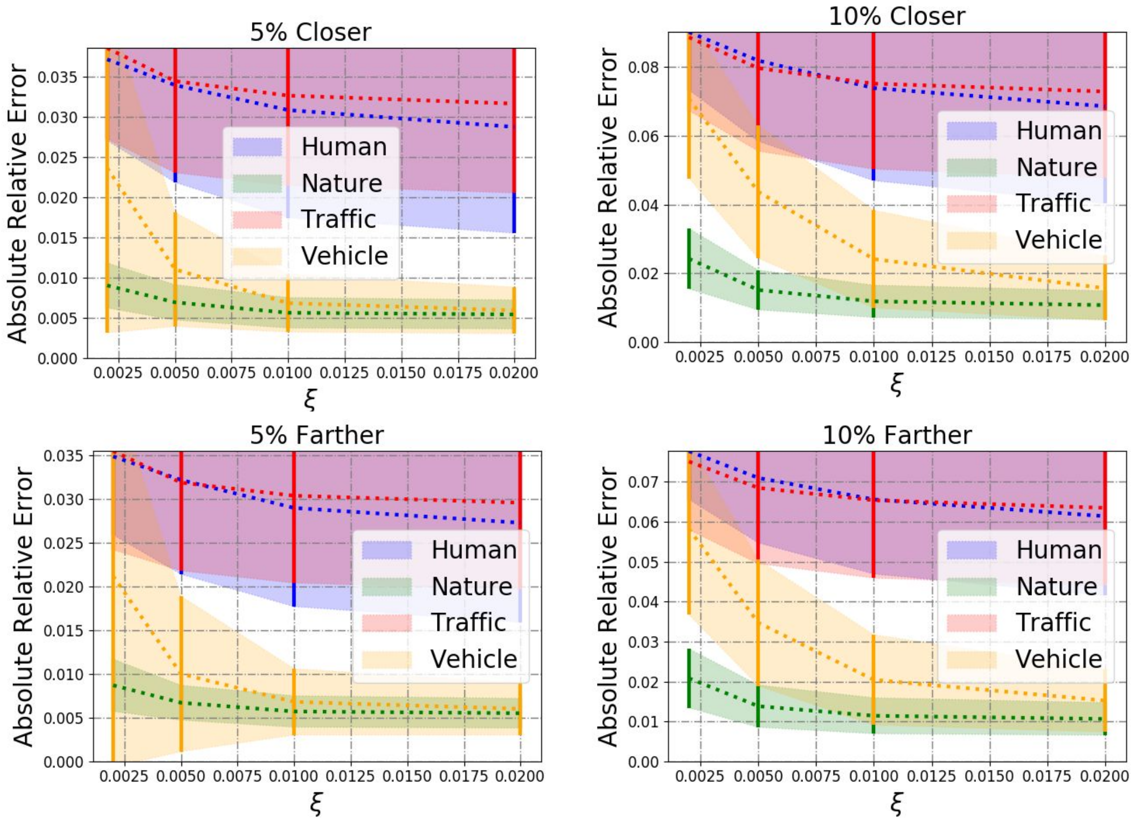


Indoor (NYU-V2)



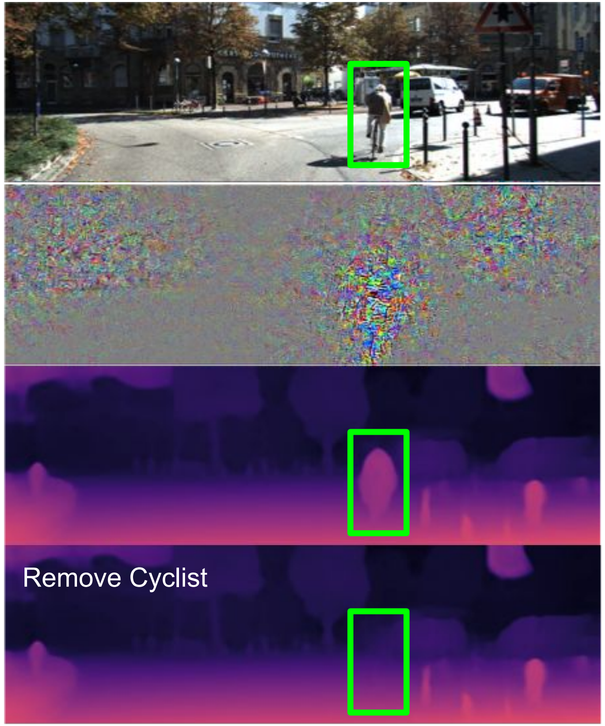
$$\text{ARE} = ||f_d(x + v(x)) - d^t(x)||_1 / d^t(x)$$

Quantitative Results
Category Conditioned Scaling

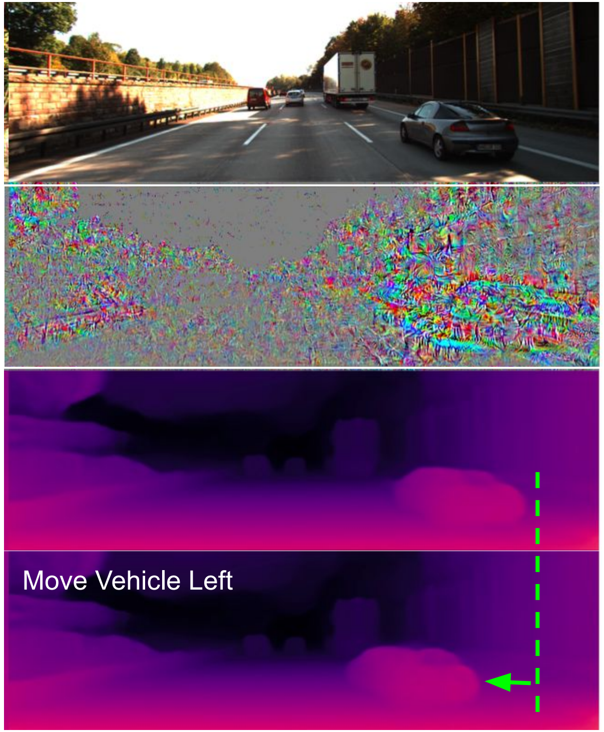


$$\text{ARE} = ||f_d(x + v(x)) - d^t(x)||_1 / d^t(x)$$

Localized Attacks on the Scene

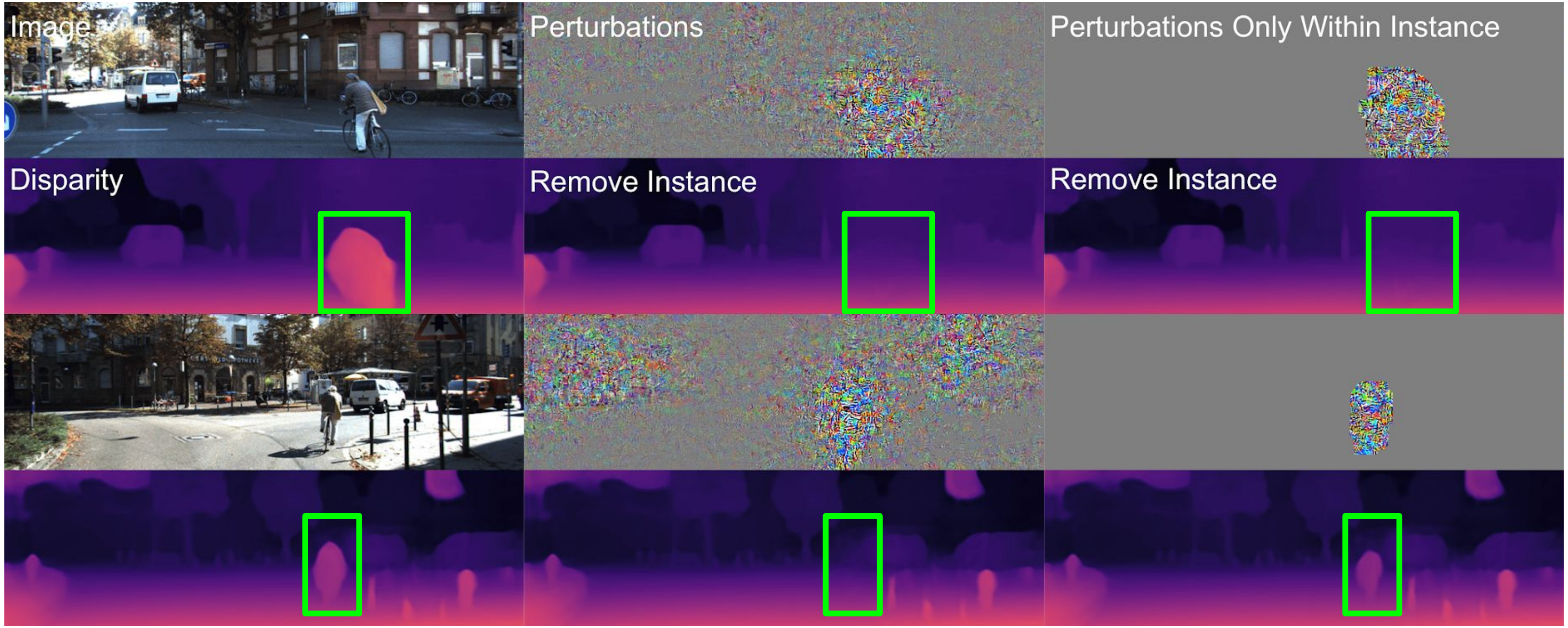


(i) removing specific instances from the scene

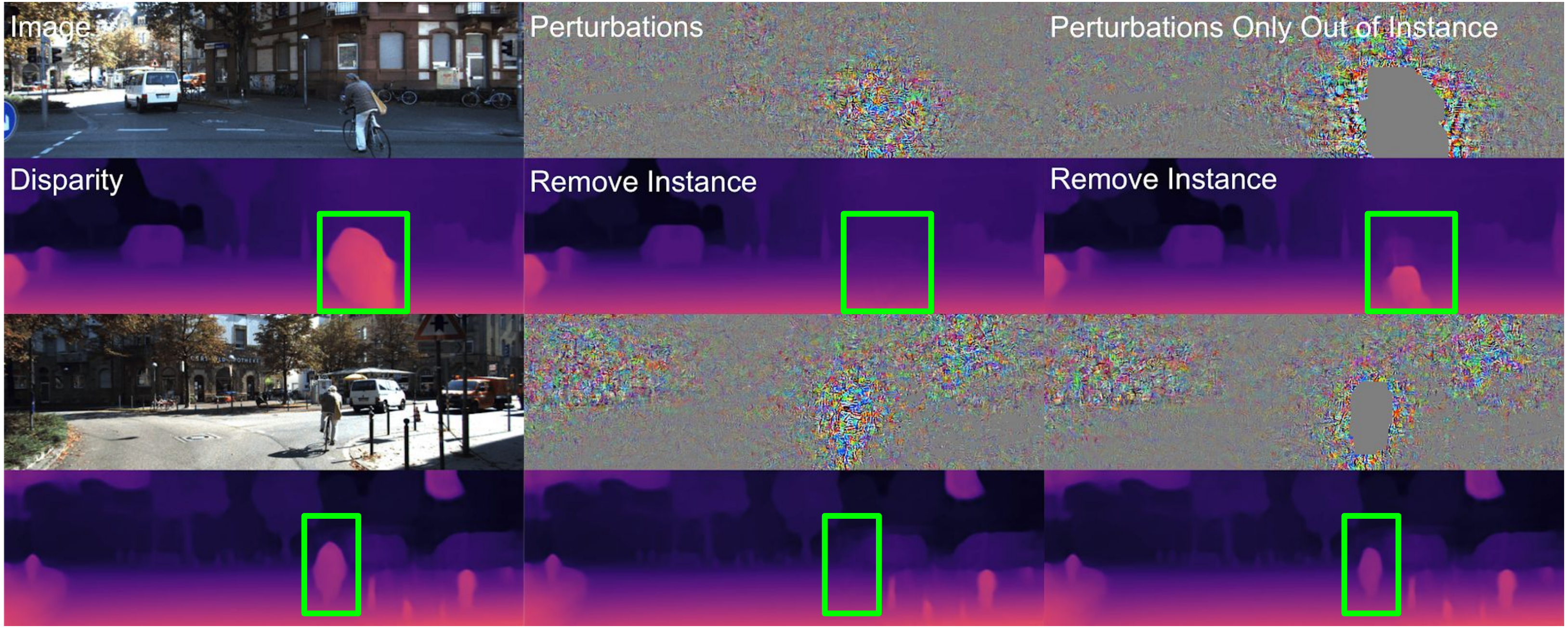


(ii) moving specific instances to different regions of the scene

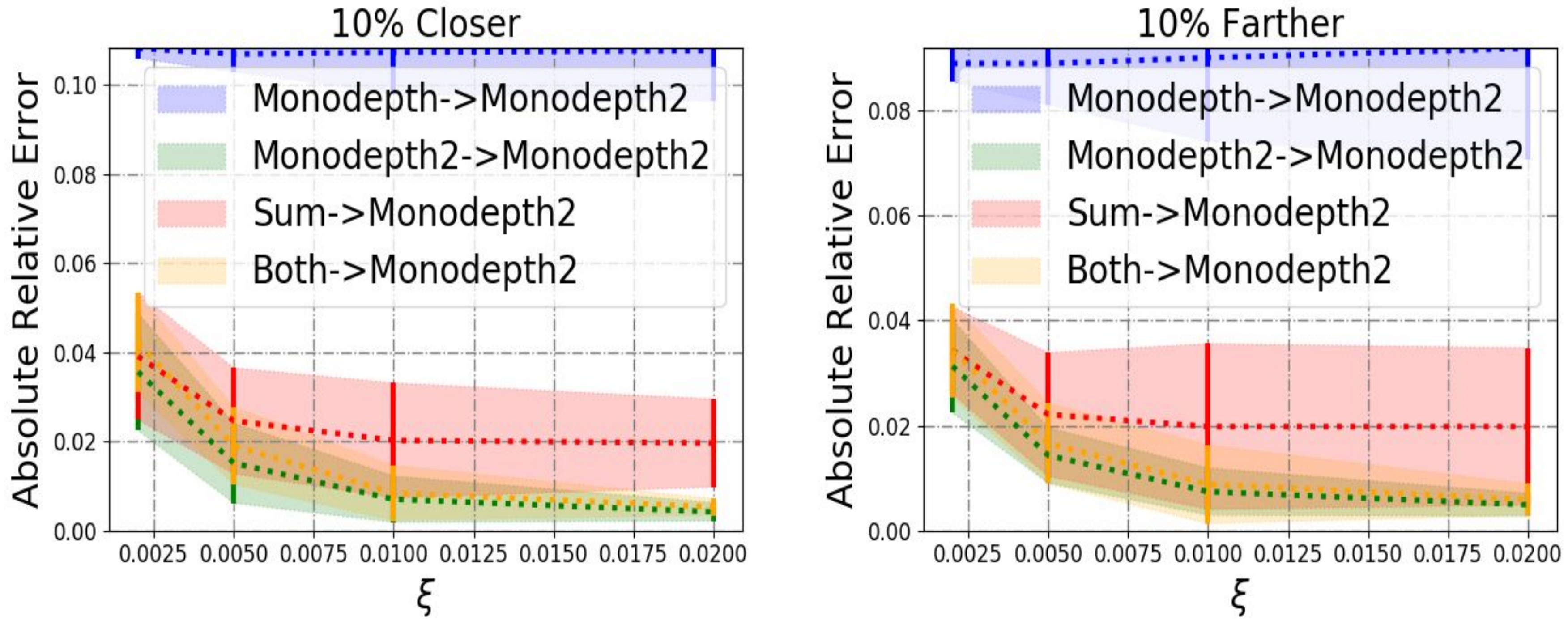
Instance Conditioned Removing



Instance Conditioned Removing



Transferability



- Fool Monodepth2 [1] with perturbations from Monodepth [2]

[1] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. ICCV 2019.
[2] C. Godard, O. Mac Aodha, G. J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR 2017.

Concluding Remarks

- SSL-semantic:
  - The proposed *speed of training* criterion shows promising results.
  - We merge the literature branched off into two different groups by smoothing on both input and weight spaces.
  - But, requirement of having real, labeled training samples for each class is not scalable.

Concluding Remarks

- SSL-semantic:
  - The proposed *speed of training* criterion shows promising results.
  - We merge the literature branched off into two different groups by smoothing on both input and weight spaces.
  - But, requirement of having real, labeled training samples for each class is not scalable.
- UDA-semantic:
  - The proposed conditional domain alignment method working well for the classification task, does not perform as well in the segmentation task.

Concluding Remarks

- SSL-semantic:
  - The proposed *speed of training* criterion shows promising results.
  - We merge the literature branched off into two different groups by smoothing on both input and weight spaces.
  - But, requirement of having real, labeled training samples for each class is not scalable.
- UDA-semantic:
  - The proposed conditional domain alignment method working well for the classification task, does not perform as well in the segmentation task.
- UDA-geometry:
  - It is possible to learn dense topology from sparse point clouds only.
  - But, it is sensitive to the density level of the input so we have to reconcile it with the image.

Concluding Remarks

- SSL-semantic:
  - The proposed *speed of training* criterion shows promising results.
  - We merge the literature branched off into two different groups by smoothing on both input and weight spaces.
  - But, requirement of having real, labeled training samples for each class is not scalable.
- UDA-semantic:
  - The proposed conditional domain alignment method working well for the classification task, does not perform as well in the segmentation task.
- UDA-geometry:
  - It is possible to learn dense topology from sparse point clouds only.
  - But, it is sensitive to the density level of the input so we have to reconcile it with the image.
- Adversarial Robustness of Unsupervised Models:
  - We show networks are vulnerable to targeted adversarial perturbations -- even to non-local ones.
  - These perturbations may not cause harm in a practical transportation application.
  - *The existence of adversaries is an opportunity.*

Acknowledgments

- My advisor, Prof. Stefano Soatto.

- Committee, Prof. Lieven Vandenberghe, Prof. Paulo Tabuada, Prof. Guy Van den Broeck.

- Collaborators,
  - Alex Wong from UCLA Vision Lab.
  - Alhussein Fawzi from Google Deepmind.
  - Ning Xu, Zhaowen Wang and Hailin Jin from Adobe Research.