北京大学
PEKING UNIVERSITY

Unsupervised Surgical Instrument Segmentation via Anchor Generation and Semantic Diffusion

Daochang Liu*, Yuhui Wei*, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li

北京大学
PEKING UNIVERSITY

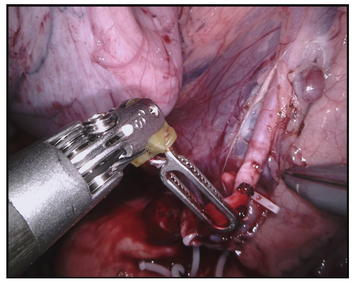
北京大学肿瘤医院
BEIJING CANCER HOSPITAL

NELVT

深睿医疗
deepwise

Introduction

Instrument Segmentation
In Surgical Video

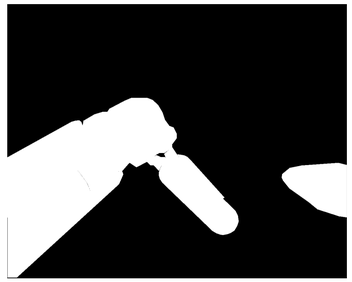


U-Net

Introduction

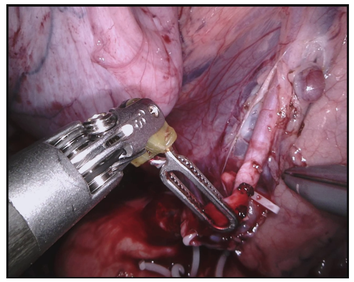
Existing Fully Supervised Methods

Very Expensive



Manual Annotation

Instrument Segmentation In Surgical Video

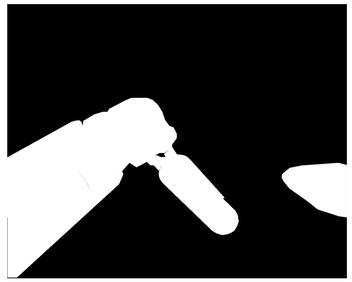


U-Net

Introduction

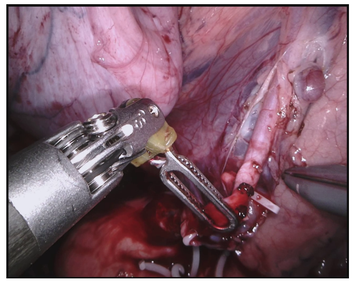
Existing Fully Supervised Methods

Very Expensive



Manual Annotation

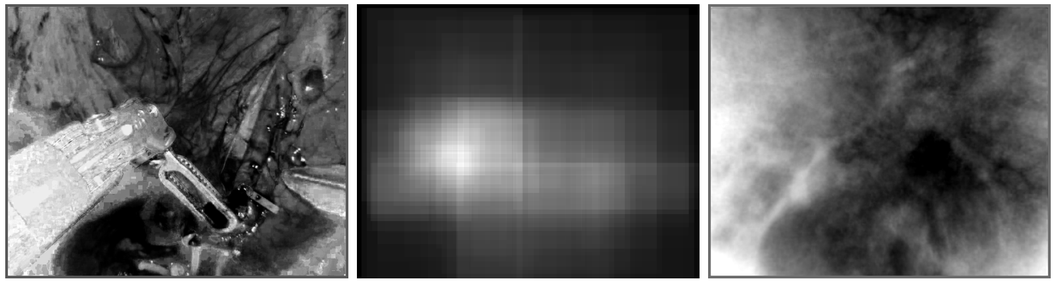
Instrument Segmentation In Surgical Video



U-Net

Our Proposed Unsupervised Method

More Affordable



Anchor Generation

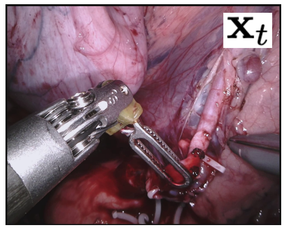
Coarse Cues (Prior Knowledge)

Semantic Diffusion

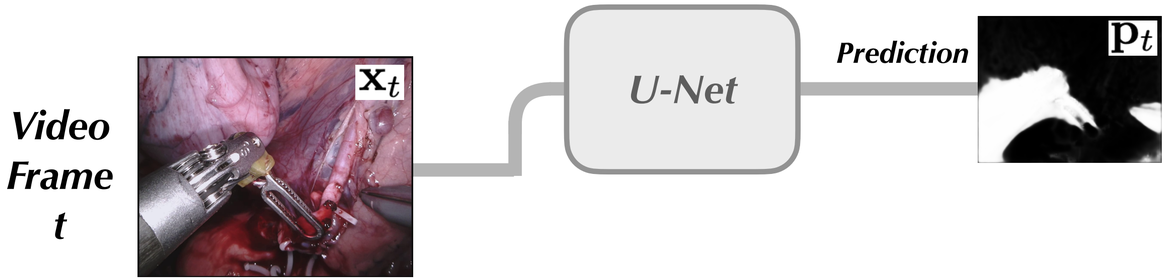
Inter-image Correlation

Method

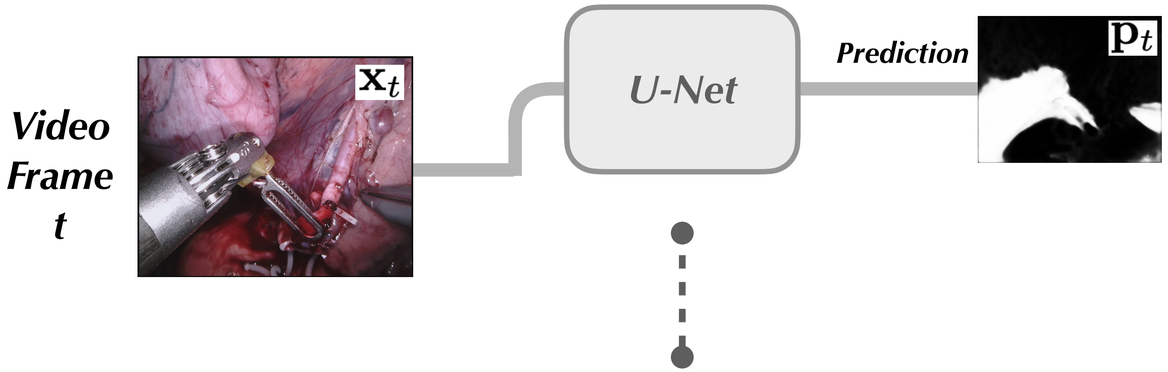
Video Frame t



Method

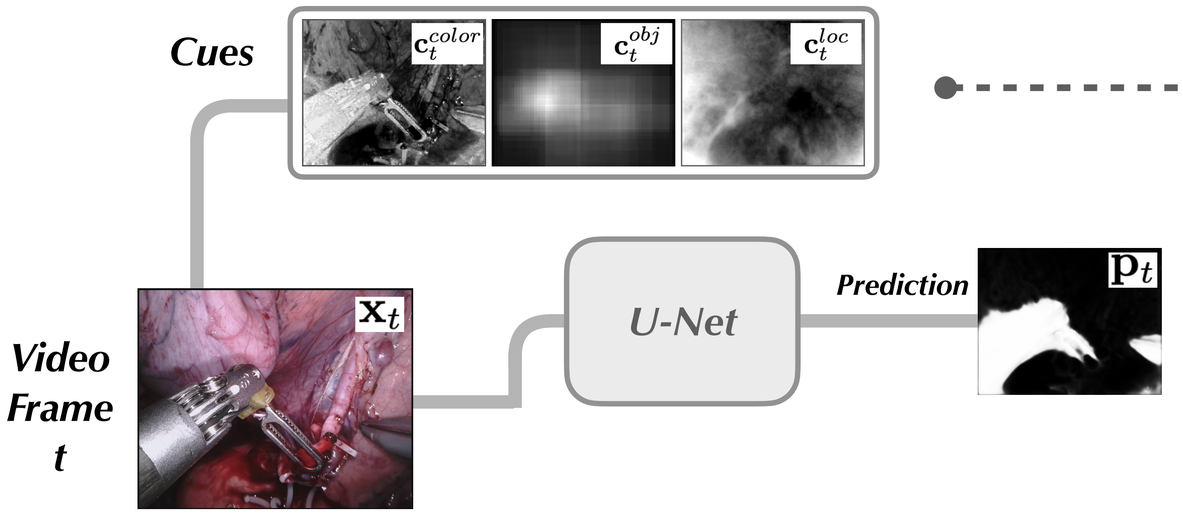


Method



How to provide training supervision without human annotation?

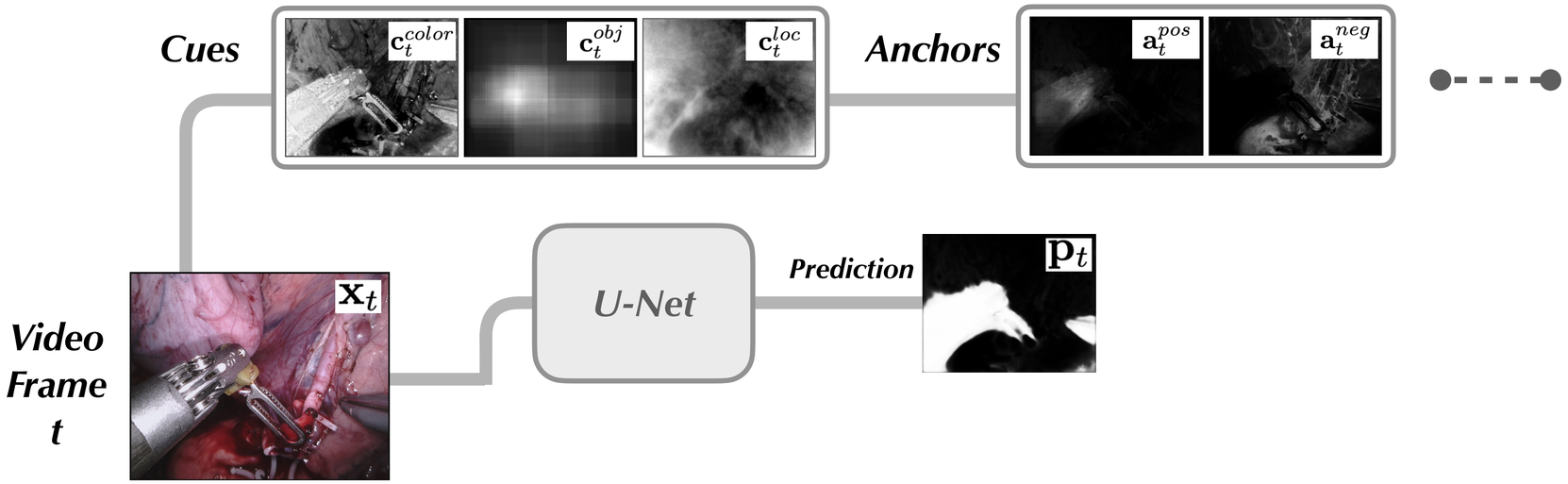
Method



Encode the knowledge about surgical instruments using coarse cues:

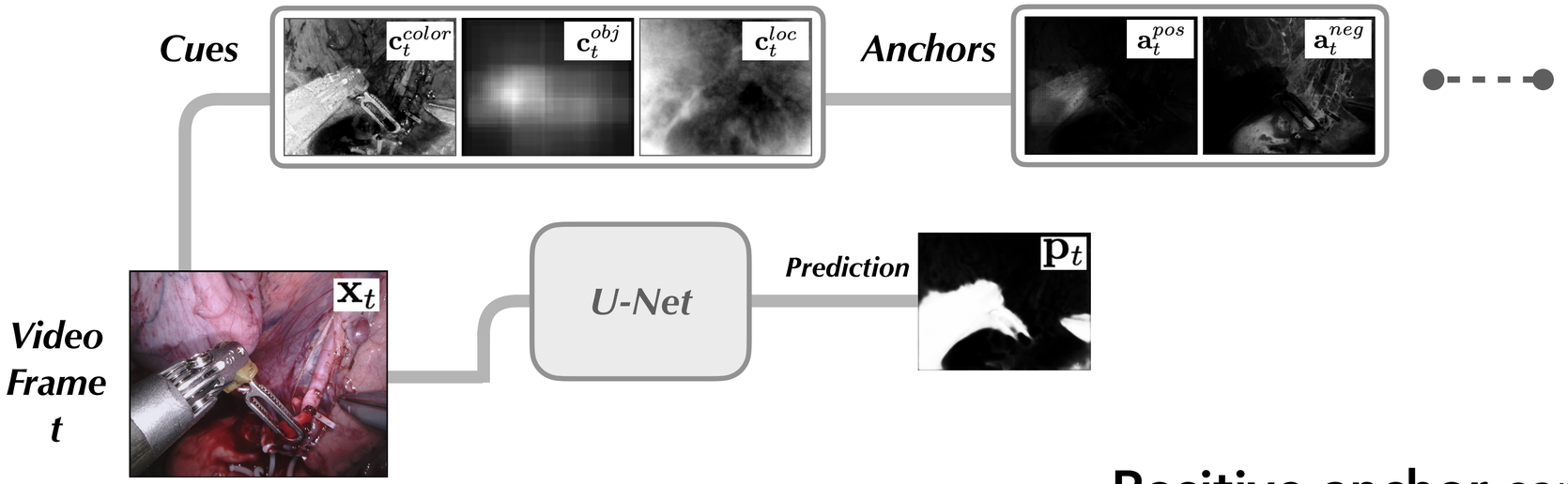
Color
Objectness
Location

Method



Generate anchors by fusing coarse cues to suppress the noise

Method



Generate anchors by fusing coarse cues to suppress the noise

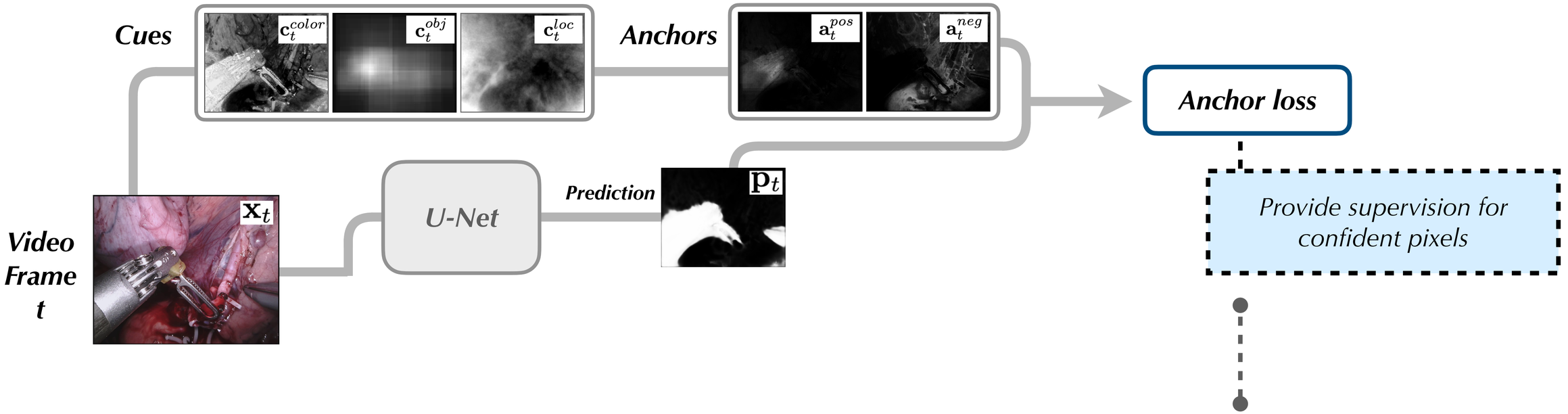
Positive anchor captures the confident instrument regions that satisfy all the cues

$$\mathbf{a}_t^{pos} = \mathbf{c}_t^{color}\mathbf{c}_t^{obj}\mathbf{c}_t^{loc}$$

Negative anchor captures the confident background regions that satisfy none of the cues

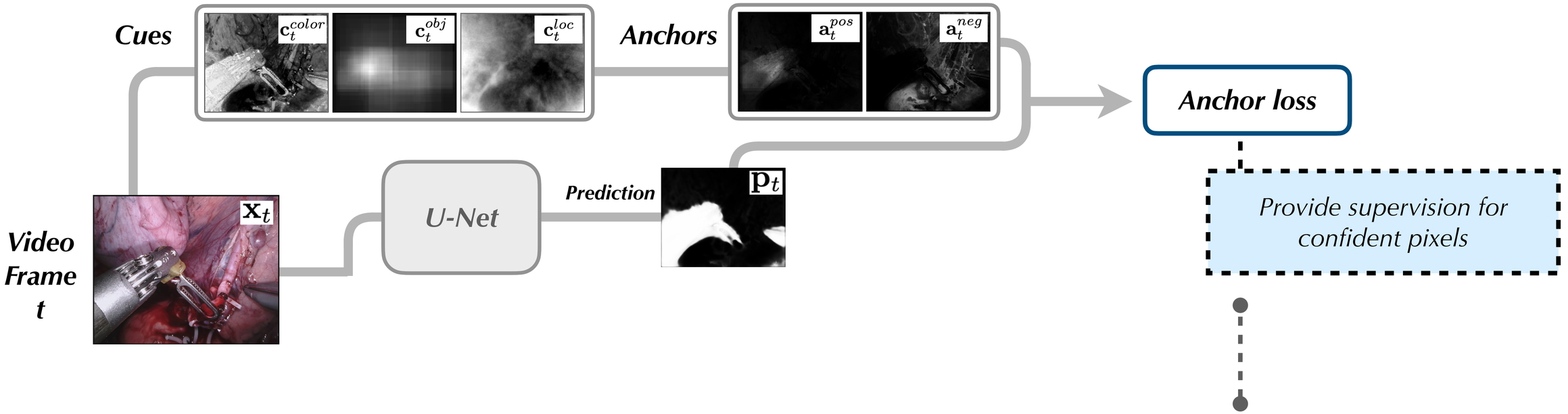
$$\mathbf{a}_t^{neg} = (\mathbf{1} - \mathbf{c}_t^{color})(\mathbf{1} - \mathbf{c}_t^{obj})(\mathbf{1} - \mathbf{c}_t^{loc})$$

Method



Anchors are used as pseudo labels to supervise confident pixels

Method

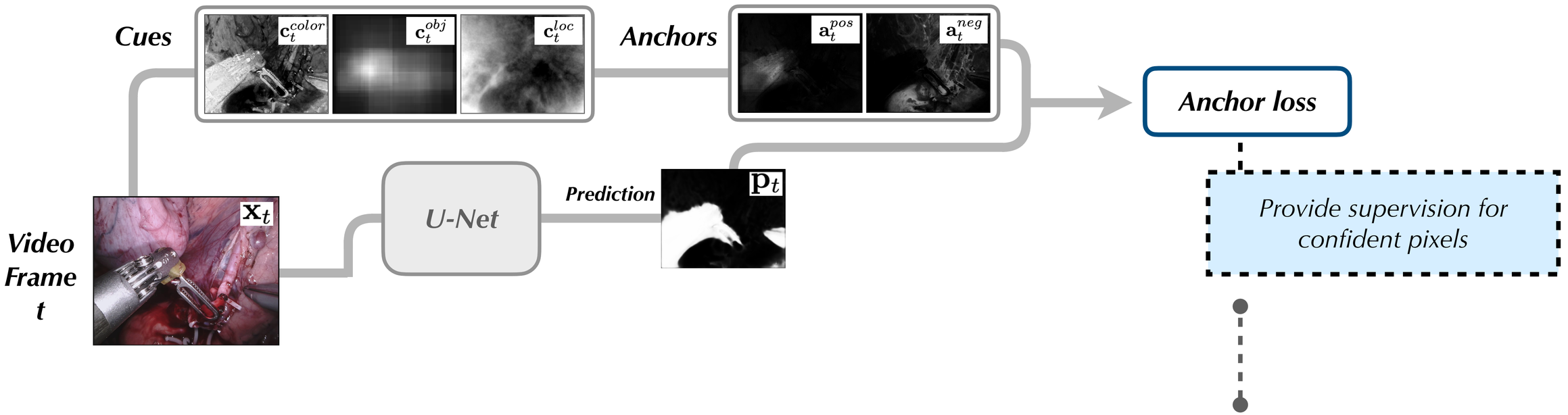


Anchor loss

$$\mathcal{L}_{anc}(\mathbf{x}_t) = \frac{1}{HW} \sum_i -\mathbf{a}_{t,i}^{pos} \mathbf{p}_{t,i} - \mathbf{a}_{t,i}^{neg}(1 - \mathbf{p}_{t,i})$$

Anchors are used as pseudo labels to supervise confident pixels

Method



Anchor loss

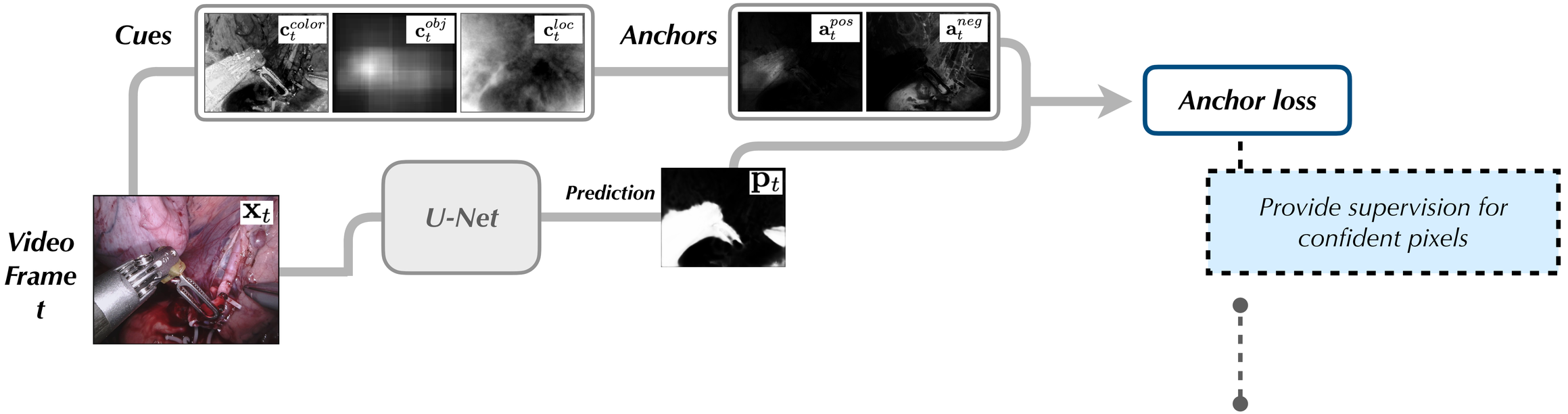
$$\mathcal{L}_{anc}(\mathbf{x}_t) = \frac{1}{HW}\sum_i -\mathbf{a}_{t,i}^{pos}\mathbf{p}_{t,i} - \mathbf{a}_{t,i}^{neg}(1-\mathbf{p}_{t,i})$$

Anchors are used as pseudo labels to supervise confident pixels

Inhibit activation on the negative anchor

Encourage activation on the positive anchor

Method



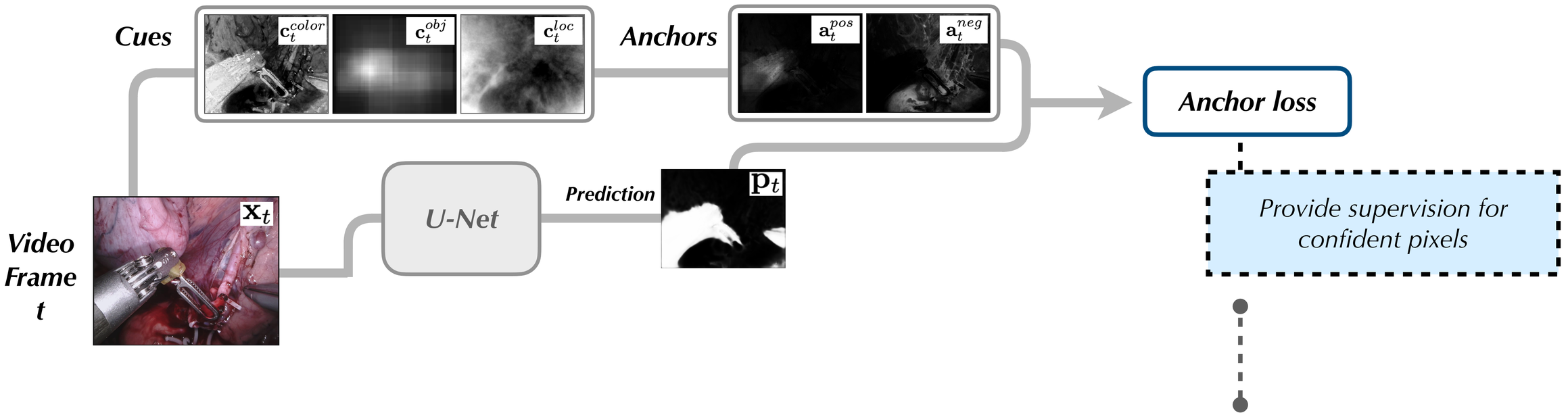
Anchor loss

$$\mathcal{L}_{anc}(\mathbf{x}_t) = \frac{1}{HW} \sum_i -\mathbf{a}_{t,i}^{pos} \mathbf{p}_{t,i} - \mathbf{a}_{t,i}^{neg}(1 - \mathbf{p}_{t,i})$$

Anchors are used as pseudo labels to supervise confident pixels

How to supervise the remaining ambiguous pixels outside the anchors?

Method



Anchor loss

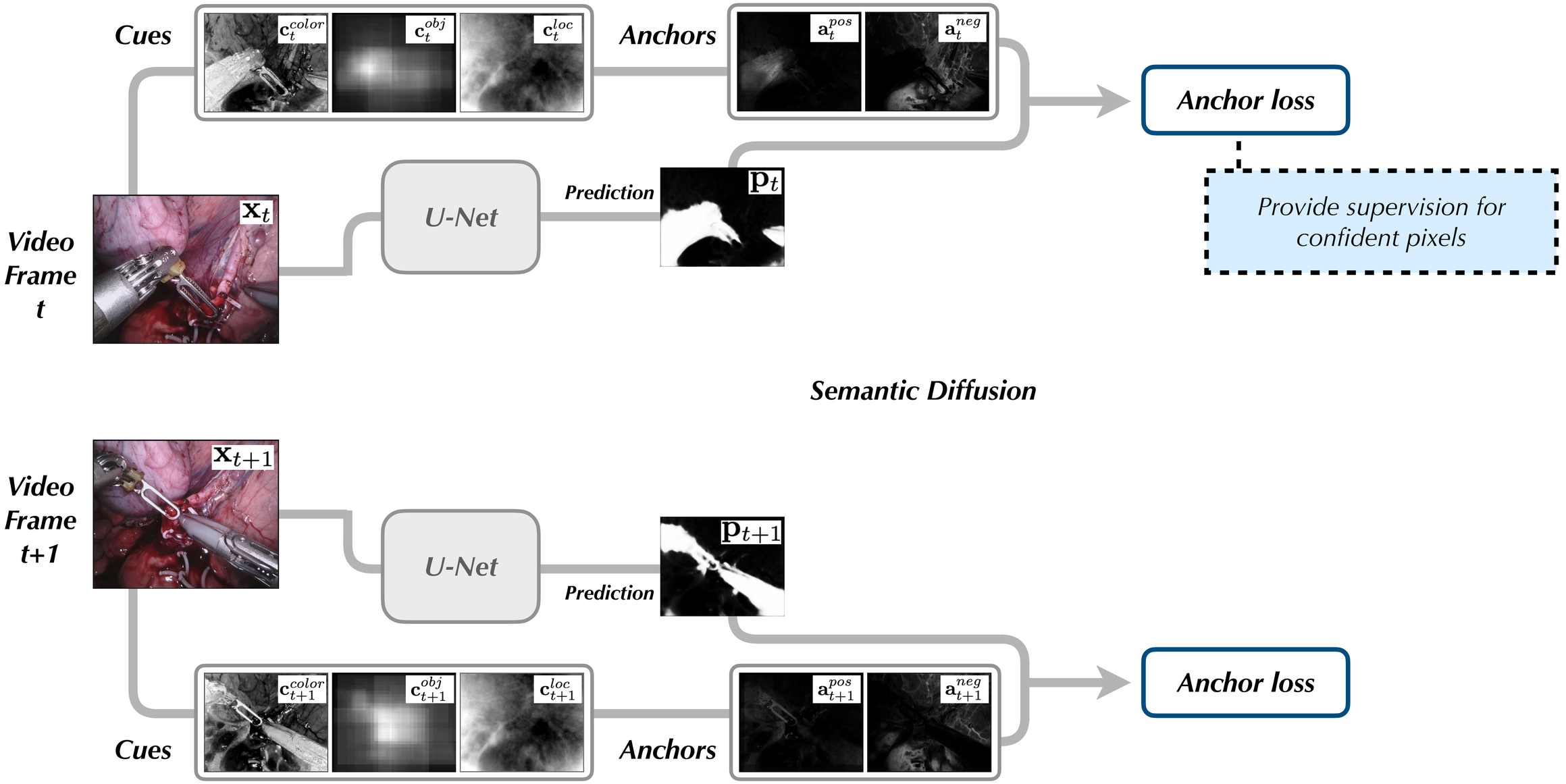
$$\mathcal{L}_{anc}(\mathbf{x}_t) = \frac{1}{HW} \sum_i -\mathbf{a}_{t,i}^{pos} \mathbf{p}_{t,i} - \mathbf{a}_{t,i}^{neg}(1 - \mathbf{p}_{t,i})$$

Anchors are used as pseudo labels to supervise confident pixels

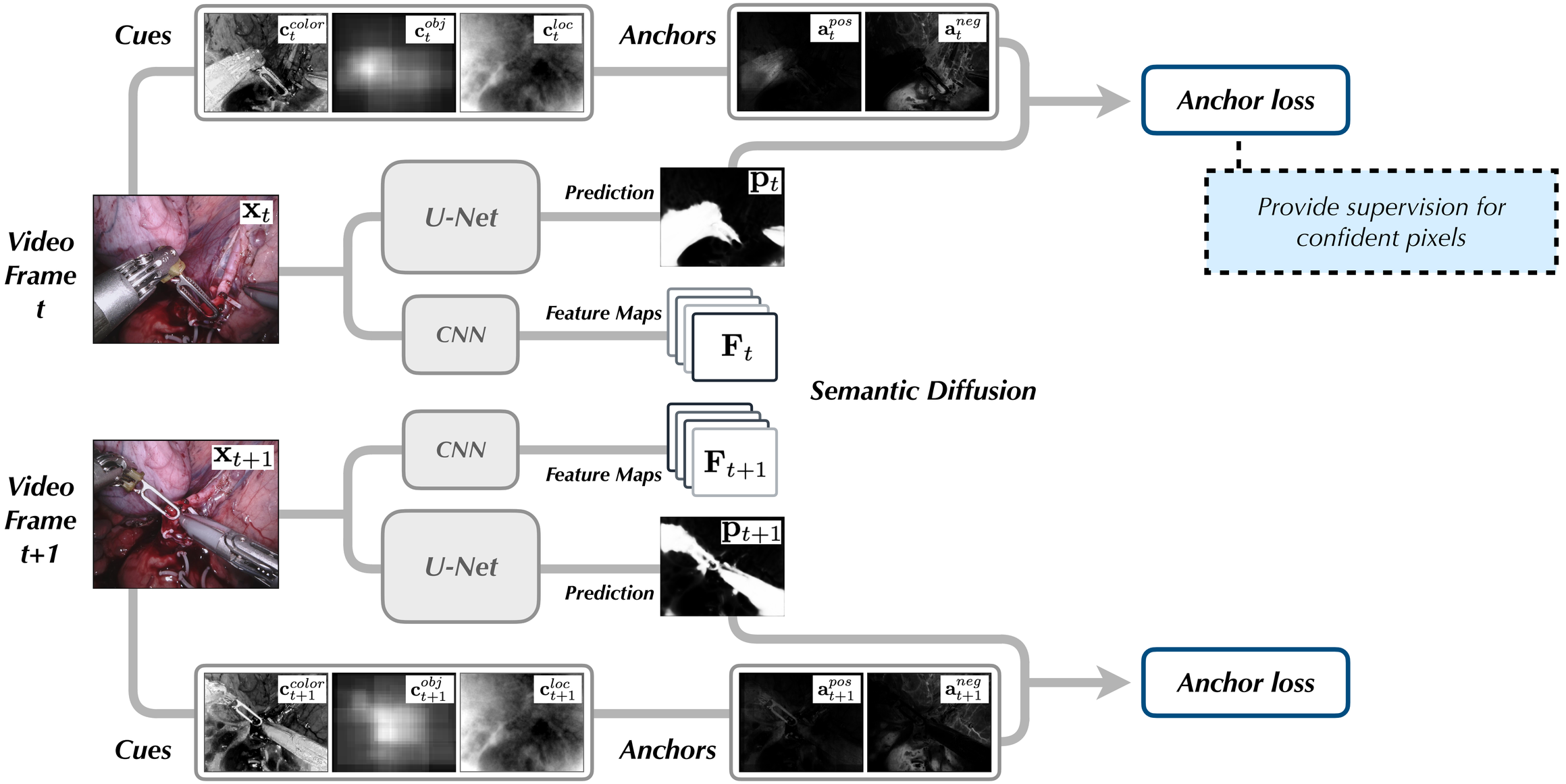
How to supervise the remaining ambiguous pixels outside the anchors?

Semantic Diffusion

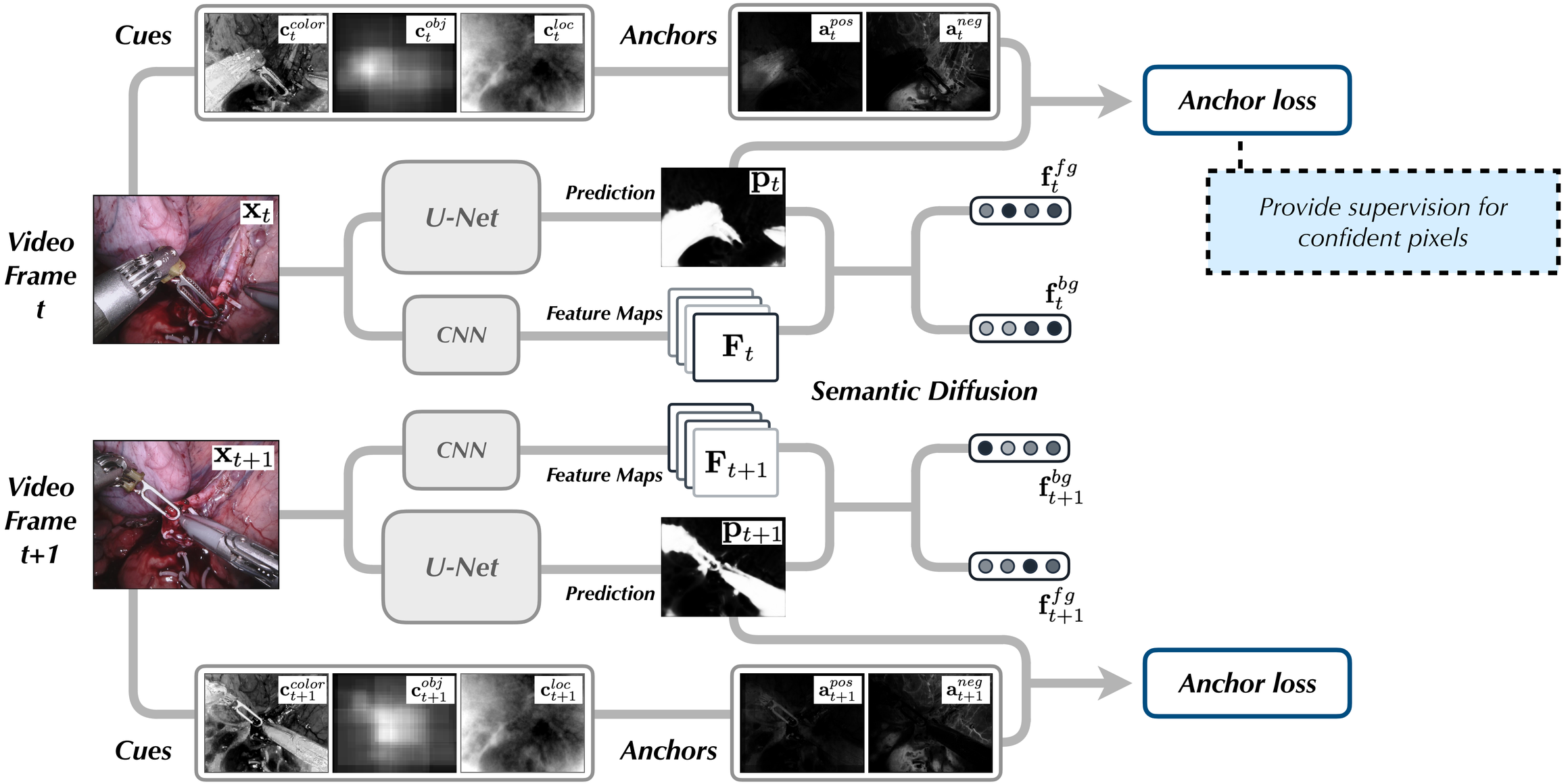
Method


Cues
$\mathbf{c}_t^{color}$
$\mathbf{c}_t^{obj}$
$\mathbf{c}_t^{loc}$
Anchors
$\mathbf{a}_t^{pos}$
$\mathbf{a}_t^{neg}$
Anchor loss
Provide supervision for confident pixels
Video Frame t
$\mathbf{x}_t$
U-Net
Prediction
$\mathbf{p}_t$
Semantic Diffusion
Video Frame t+1
$\mathbf{x}_{t+1}$
U-Net
Prediction
$\mathbf{p}_{t+1}$
Anchor loss
Cues
$\mathbf{c}_{t+1}^{color}$
$\mathbf{c}_{t+1}^{obj}$
$\mathbf{c}_{t+1}^{loc}$
Anchors
$\mathbf{a}_{t+1}^{pos}$
$\mathbf{a}_{t+1}^{neg}$

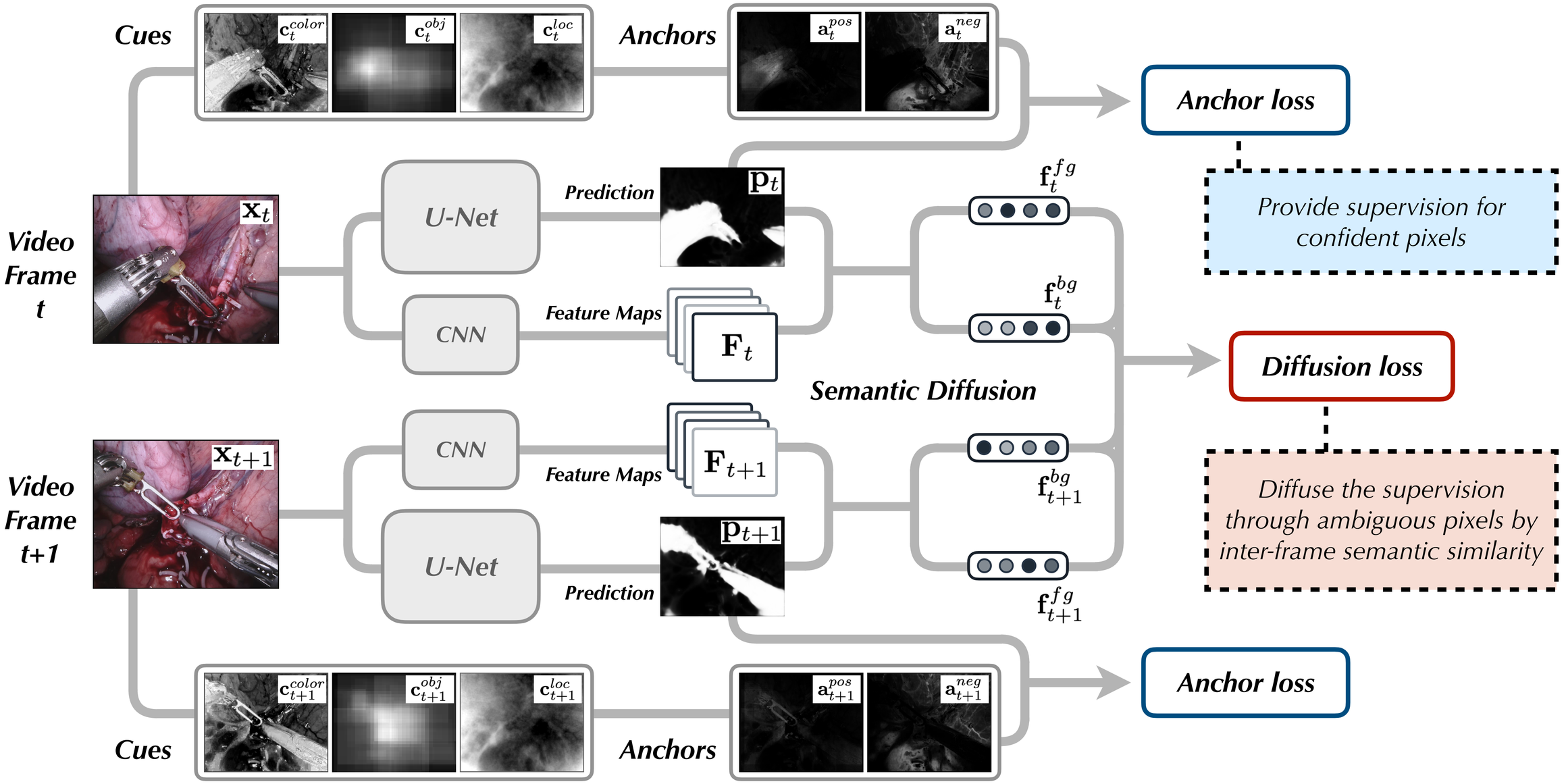
Method


Cues
$\mathbf{c}_t^{color}$
$\mathbf{c}_t^{obj}$
$\mathbf{c}_t^{loc}$
Anchors
$\mathbf{a}_t^{pos}$
$\mathbf{a}_t^{neg}$
Anchor loss
Provide supervision for confident pixels
Video Frame t
$\mathbf{x}_t$
U-Net
Prediction
$\mathbf{p}_t$
CNN
Feature Maps
$\mathbf{F}_t$
Semantic Diffusion
Video Frame t+1
$\mathbf{x}_{t+1}$
CNN
Feature Maps
$\mathbf{F}_{t+1}$
U-Net
Prediction
$\mathbf{p}_{t+1}$
Cues
$\mathbf{c}_{t+1}^{color}$
$\mathbf{c}_{t+1}^{obj}$
$\mathbf{c}_{t+1}^{loc}$
Anchors
$\mathbf{a}_{t+1}^{pos}$
$\mathbf{a}_{t+1}^{neg}$
Anchor loss

Method


Cues
$\mathbf{c}_t^{color}$
$\mathbf{c}_t^{obj}$
$\mathbf{c}_t^{loc}$
Anchors
$\mathbf{a}_t^{pos}$
$\mathbf{a}_t^{neg}$
Anchor loss
Provide supervision for confident pixels
Video Frame t
$\mathbf{x}_t$
U-Net
Prediction
$\mathbf{p}_t$
CNN
Feature Maps
$\mathbf{F}_t$
$\mathbf{f}_t^{fg}$
$\mathbf{f}_t^{bg}$
Semantic Diffusion
Video Frame t+1
$\mathbf{x}_{t+1}$
CNN
Feature Maps
$\mathbf{F}_{t+1}$
U-Net
Prediction
$\mathbf{p}_{t+1}$
$\mathbf{f}_{t+1}^{bg}$
$\mathbf{f}_{t+1}^{fg}$
Cues
$\mathbf{c}_{t+1}^{color}$
$\mathbf{c}_{t+1}^{obj}$
$\mathbf{c}_{t+1}^{loc}$
Anchors
$\mathbf{a}_{t+1}^{pos}$
$\mathbf{a}_{t+1}^{neg}$
Anchor loss

Method


Cues
Anchors
Anchor loss
Provide supervision for confident pixels
Video Frame t
U-Net
Prediction
CNN
Feature Maps
Semantic Diffusion
Diffusion loss
Video Frame t+1
CNN
Feature Maps
U-Net
Prediction
Diffuse the supervision through ambiguous pixels by inter-frame semantic similarity
Cues
Anchors
Anchor loss

Method

Diffusion loss

$$\mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{fg}, \mathbf{f}_{t+1}^{fg}) + m^{fg}, 0)$$

Method

Diffusion loss

$$\mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{fg}, \mathbf{f}_{t+1}^{fg}) + m^{fg}, 0)$$

Adjacent video frames

Semantic features of the foreground and the background regions in the two frames

Margin hyperparameter

Method

Diffusion loss

$$\mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{fg}, \mathbf{f}_{t+1}^{fg}) + m^{fg}, 0)$$

Maximize inter-frame instrument-instrument similarity

Minimize intra-frame instrument-background similarities

Method

Diffusion loss

$$\mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{fg}, \mathbf{f}_{t+1}^{fg}) + m^{fg}, 0)$$

$$\mathcal{L}_{dif}^{bg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{bg}, \mathbf{f}_{t+1}^{bg}) + m^{bg}, 0)$$

Similarly, maximize the inter-frame background-background similarity

Method

Diffusion loss

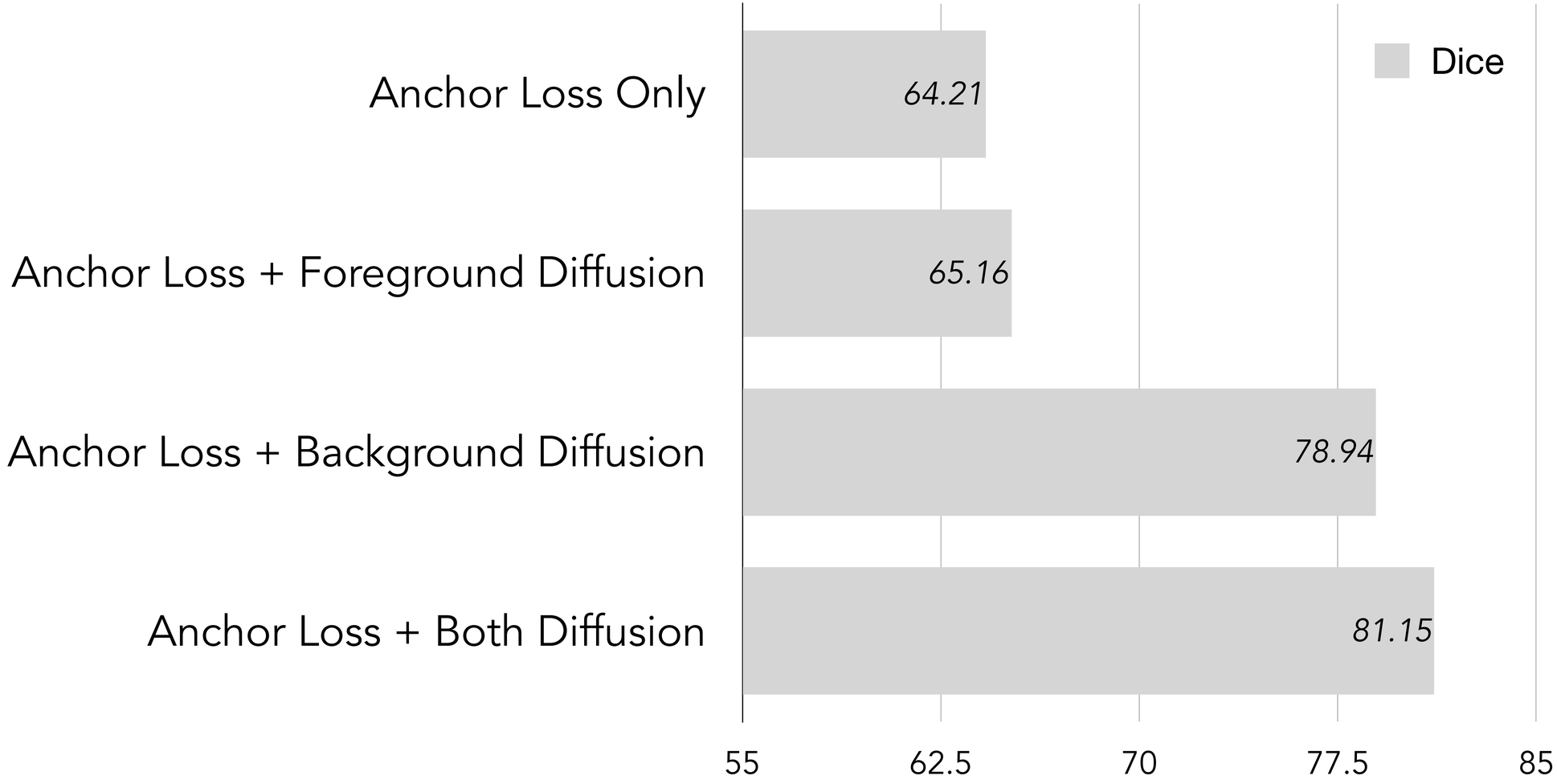
$$\mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{fg}, \mathbf{f}_{t+1}^{fg}) + m^{fg}, 0)$$

$$\mathcal{L}_{dif}^{bg}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \max(\phi(\mathbf{f}_t^{fg}, \mathbf{f}_t^{bg}) + \phi(\mathbf{f}_{t+1}^{fg}, \mathbf{f}_{t+1}^{bg}) - 2\phi(\mathbf{f}_t^{bg}, \mathbf{f}_{t+1}^{bg}) + m^{bg}, 0)$$

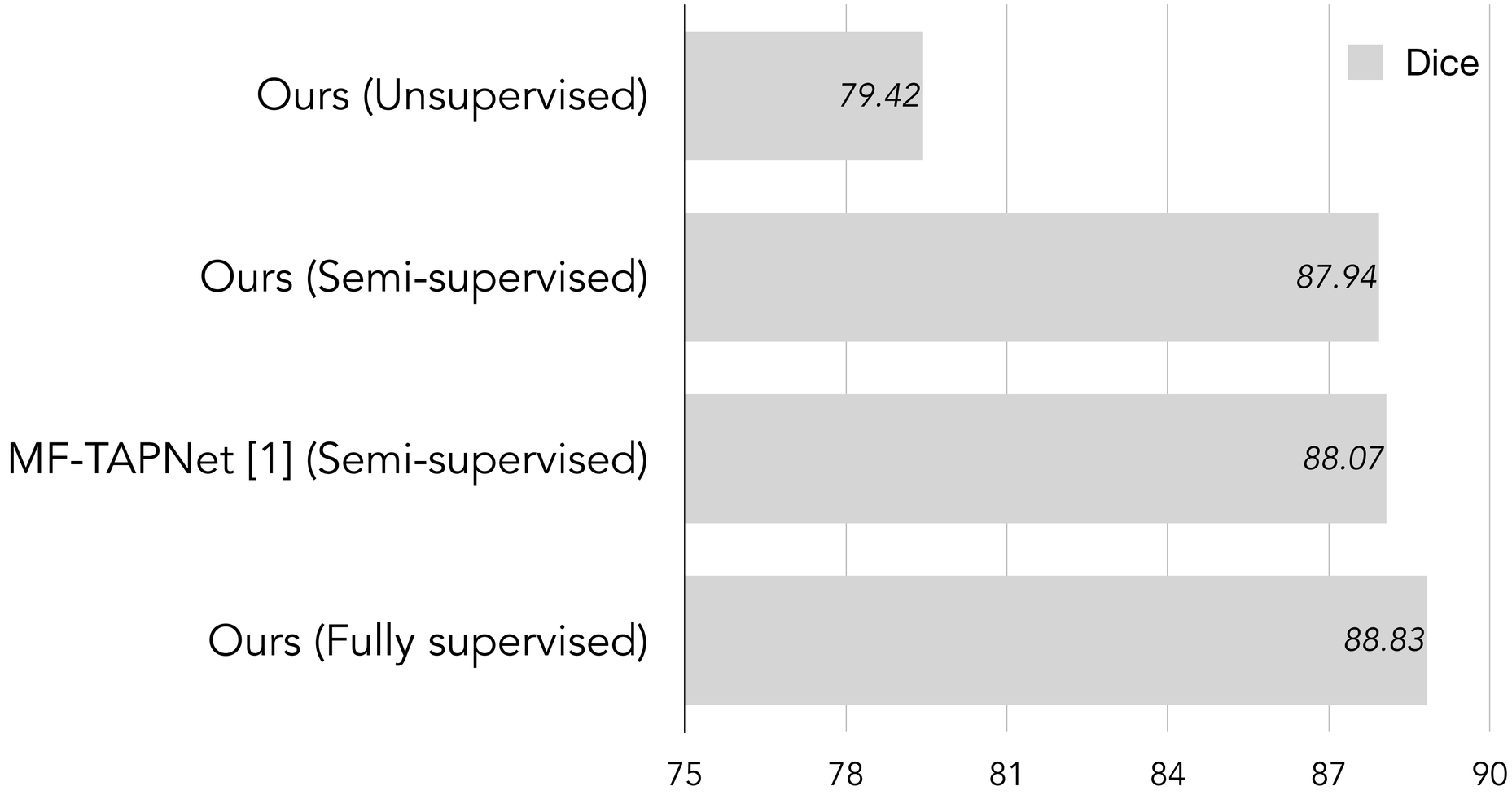
Full loss = Anchor loss + Diffusion loss

$$\mathcal{L}_{full}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \mathcal{L}_{anc}(\mathbf{x}_t) + \mathcal{L}_{anc}(\mathbf{x}_{t+1}) + \mathcal{L}_{dif}^{fg}(\mathbf{x}_t, \mathbf{x}_{t+1}) + \mathcal{L}_{dif}^{bg}(\mathbf{x}_t, \mathbf{x}_{t+1})$$

Results on EndoVis 2017



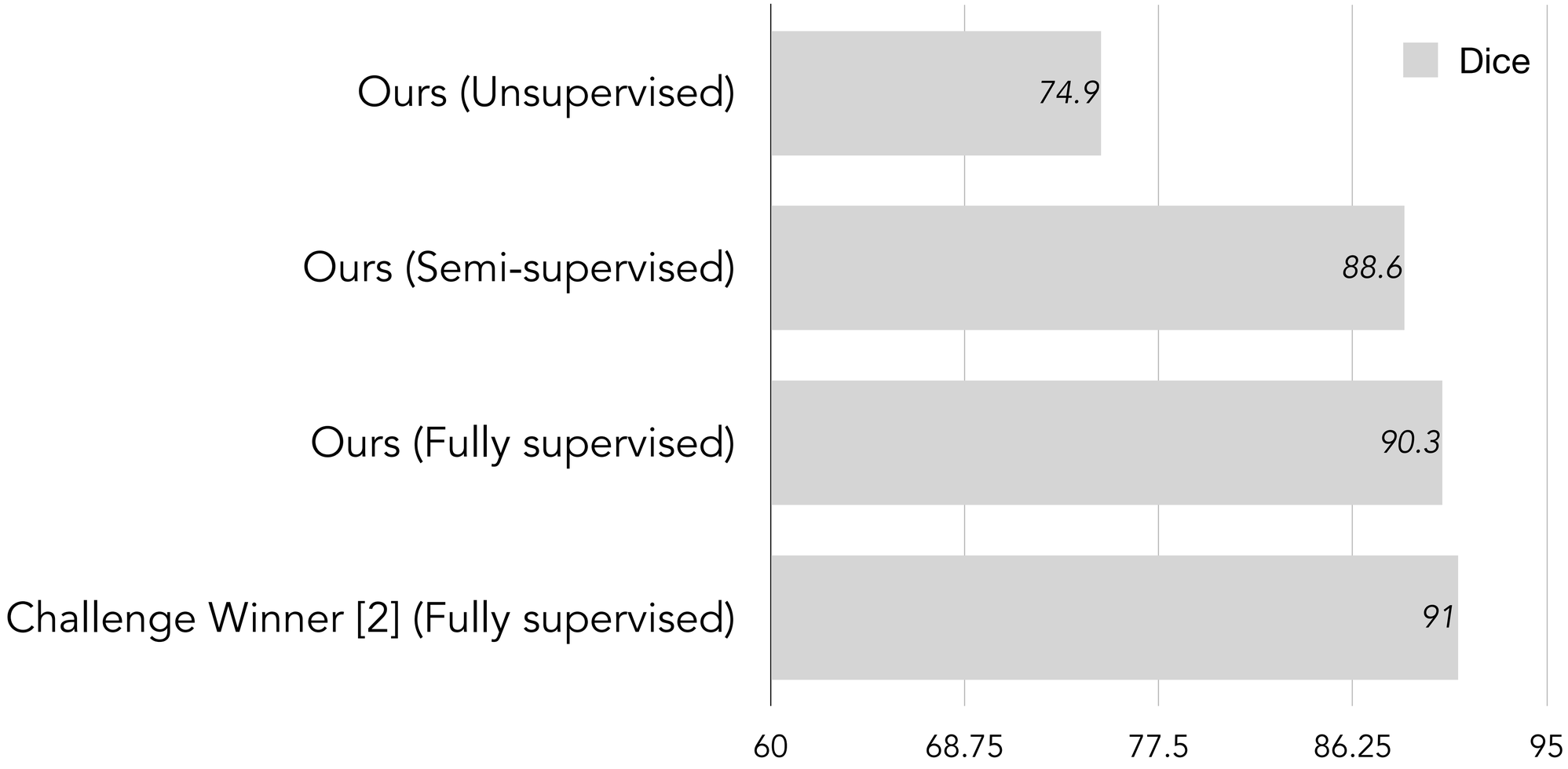
Extend to Semi-Supervised and Fully Supervised Settings



We only use vanilla U-Net

[1] Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: MICCAI (2019)

Extend to Skin Lesion Segmentation on ISIC 2016



We only use vanilla U-Net

[2] Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the inter- national skin imaging collaboration (ISIC). arXiv:1605.01397 (2016)

Video Demo

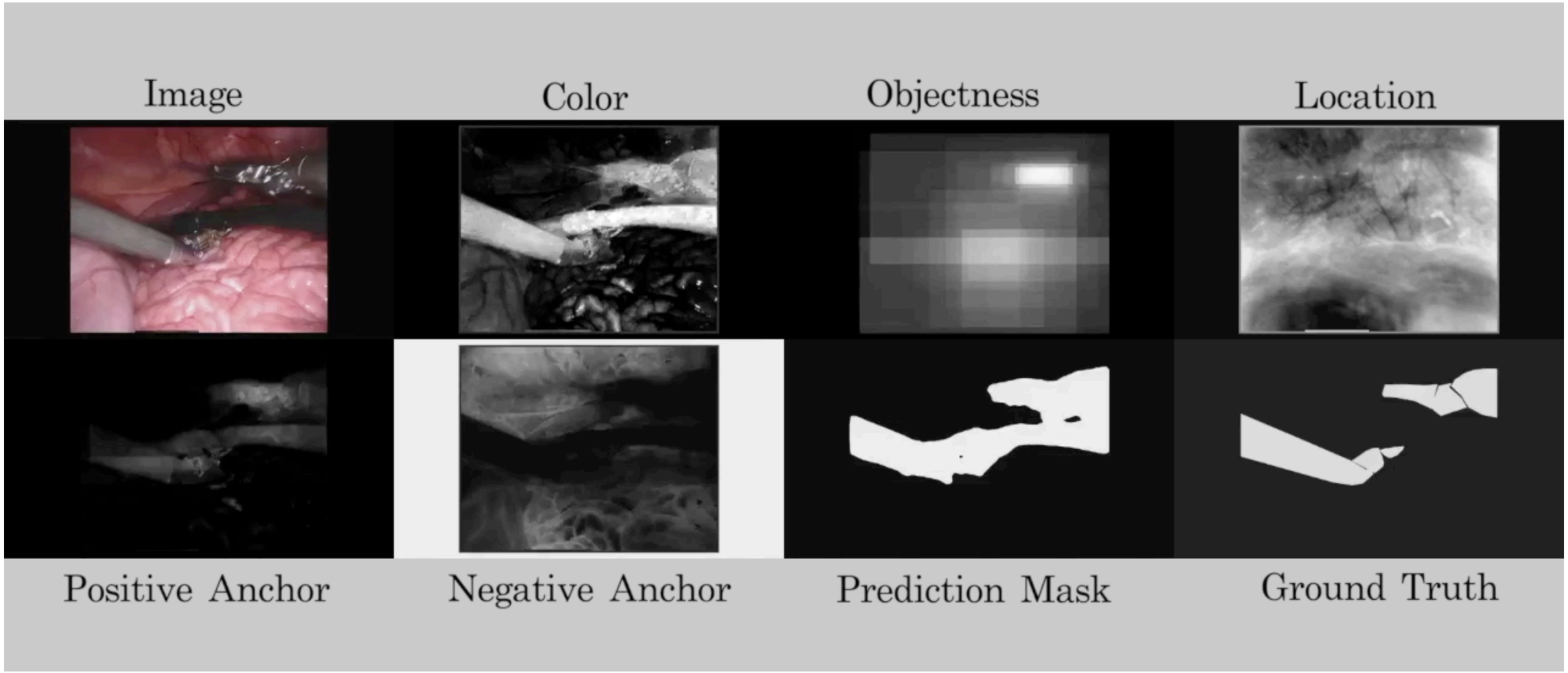


Image
Color
Objectness
Location
Positive Anchor
Negative Anchor
Prediction Mask
Ground Truth

Conclusion

- **Unsupervised** Surgical Instrument Segmentation
- No human annotation needed
- **Anchor generation** to supervise confident pixels
- **Semantic diffusion** to supervise ambiguous pixels
- Promising results on **EndoVis 2017** and **ISIC 2016**
- **Codes** at https://github.com/Finspire13/AGSD-Surgical-Instrument-Segmentation

Thank you!

Welcome to Discuss Online.

ORAL SESSION WEDS-AM-4
6:30 PM - 7:00 PM CST on Wednesday, 7 October

POSTER SESSION 5
5:00 PM - 6:30 PM CST on Wednesday, 7 October